Tampere University

Aowen Shi

# APPLICATION OF TRANSFORMER NEURAL NETWORKS TO EEG SIGNAL ANALYSIS

Master of Science Thesis
Faculty of Information Technology and Communication Sciences
Examiner: Prof. Tarmo Lipping
Jari Turunen
September 2023

# ABSTRACT

Aowen Shi: Application of Transformer Neural Networks to EEG Signal Analysis
Master of Science Thesis
Tampere University
September 2023

Transformer networks have emerged as an important advancement in the field of deep learning and are widely used in several contemporary domains. Transformer networks were originally developed for natural language processing (NLP) and have shown potential for efficiently collecting complex patterns in electroencephalography (EEG) data. This thesis briefly overviews the basic ideas behind cognitive load assessment and transformer networks. The paper also reviews previous research exploring the use of transformer networks in EEG analysis. In addition, a case study is conducted to illustrate the application of transformer networks, followed by a comprehensive discussion of the results obtained.

The review of previous studies includes those that have used transformer networks alone, as well as those that have been combined with other network architectures. The review of their studies and their results shows that the transformer and its combined architectures have obtained good results in the classification task in the direction of EEG analysis. In the experimental part of the case study, the EEG conformer network was experimented with a Python environment using local data and a public dataset named simultaneous task EEG workload (STEW). The experimental results show that the training results of EEG conformer are closely related to the data complexity and the difficulty of the classification task and that the architecture of this model leads to a high demand on the amount of data and is prone to overfitting. In addition, this model is sensitive to parameter variations, and the optimal parameters for different datasets have large differences.

According to the existing research results, transformers are considered to play a crucial role in the development of deep learning. Moreover, this thesis concludes by revealing prospective challenges and issues that deserve attention in the future adoption of transformer networks. This means that transformers have more possibilities in the field of EEG analysis in the future, thus bringing more help to people in real life, such as the diagnosis of neurological diseases, sleep studies, cognitive neuroscience research, brain-computer interfaces (BCI), and so on.

Keywords: Electroencephalogram (EEG), transformer network, cognitive load, classification

The default GPT-3.5 of ChatGPT has been used to embellish and rephrase statements.

# PREFACE

This thesis discusses the use of transformers in EEG analysis and analyses an example study. The use of transformer networks in EEG analysis has the potential to advance our understanding of the human mind and benefit cognitive psychology, clinical neurology, and human-computer interaction research, among others. Future breakthroughs in EEG analysis and our comprehension of cognitive processes can be anticipated as researchers and practitioners continue to refine and expand these applications.

Throughout the completion of my dissertation, both the dissertation and the experiment, Professor Tarmo was of great assistance. During my time as a research assistant in Professor Tarmo's group, I developed a strong interest in this discipline. I asked Professor Tarmo for permission to conclude my thesis on this subject under his supervision, and he graciously agreed. During the experiment, when I encountered difficulties, Prof. Tarmo would reply to my emails very quickly and offer me some guidance, which also helped me to have a deeper understanding of this field. I would like to express my gratitude to Professor Tarmo Lipping. In addition, I would like to thank Jari Turunen for serving as my second examiner.

In addition, I would like to thank my parents for their financial and mental support of my studies, as well as their early education, which allowed me to earn my graduate degree step by step.

Last but not least, I'd like to acknowledge my significant other, Jinhan, for his affection and support, and for comforting and encouraging me during a time of difficulty. I hope that we will continue to be able to comprehend each other so well and provide each other with support as we continue our journey!

Xianning, 25th September 2023

Aowen Shi

# CONTENTS

# 1. INTRODUCTION

The transformer architecture stands as a revolutionary paradigm in the dynamic field of artificial intelligence and deep learning technologies. Tailored initially for natural language processing endeavors, its adaptability, and capacity to capture intricate interdependencies within sequences have transcended the confines of text-centric applications. This remarkable versatility has kindled widespread interest across diverse domains, including the realm of neuroscience. Here, the multifaceted dynamics of Electroencephalogram (EEG) signals demand advanced computational methodologies.

The convergence of transformer networks and EEG signal analysis presents a captivating fusion of innovative technology and the depths of cognitive science. EEG, serving as a direct reflection of neural activity, provides a microscopic portal into the cognitive machinery guiding human behavior. Notably, the assessment of cognitive load assumes a pivotal role in deciphering the allocation of cognitive resources during tasks, influencing facets ranging from decision-making to problem-solving and overall task performance. Moreover, the capacity to gauge cognitive load holds far-reaching implications, spanning the optimization of human-computer interfaces to the enhancement of pedagogical methods and medical diagnostic practices.

The thesis unfolds within a sequence of chapters, each contributing to the overarching mission of harnessing transformer networks' prowess for EEG signal analysis. The inaugural segment of the thesis embarks upon EEG-based cognitive load assessment, elucidating its fundamental significance and its pertinence to practical, real-world applications.

Subsequently, the thesis embarks on a journey into the core tenets of the transformer network, unveiling its fundamental architecture and operational mechanisms. The spotlight here is on its self-attention mechanism and depth architecture, which lay the essential theoretical groundwork, enabling a deeper comprehension of subsequent applications in EEG signal analysis.

Building upon this foundation, the literature review chapter embarks on an extensive survey of prior research endeavors that have harnessed transformer networks to dissect EEG signals. This survey assesses the strides made in leveraging this innovative neural architecture to extract valuable insights from EEG data, presenting a view of the existing corpus of knowledge.

The centerpiece of this thesis is the case study, wherein the EEG Conformer Network, an architectural marvel that marries the power of a Convolutional Neural Network (CNN) with a transformer network, as proposed by Song et al. [1], takes center stage. This amalgamation significantly amplifies the network's capability to apprehend both spatial and temporal correlations within EEG data. This section serves as a testament to the practical application of the Transformer architecture, adeptly addressing the formidable challenges that EEG analysis entails, encompassing data pre-processing techniques and network fine-tuning.

Finally, the thesis provides a thoughtful discussion of the findings and their implications in addition to drawing conclusions. The research journey of this dissertation demonstrates the potential of transformer neural networks to reveal the complex patterns embedded in EEG signals, thereby advancing our understanding of cognitive processes and providing a promising avenue for the practical application of cognitive load assessment.

# 2. ASSESSMENT OF COGNITIVE LOAD BASED ON EEG

Electroencephalography (EEG) is a valuable technique for assessing cognitive burden within the fields of cognitive neuroscience and human-computer interaction. The term "cognitive load" refers to the quantity of mental work and processing demands that are placed on the cognitive resources of an individual when they are carrying out a task. It is extremely important to do a cognitive load assessment since an excessive amount of cognitive load can result in lower performance, an increase in the number of errors made, and mental tiredness. On the other hand, a healthy amount of mental strain is beneficial to both learning and performance on tasks. The measurement of cognitive load using electroencephalography (EEG) is a useful technique that may be used in user interface design, education, and healthcare to determine the mental strain and processing demands placed on a person while completing a task.

## 2.1 Electroencephalography

Recording electrical activity produced by the brain using the EEG is a non-invasive procedure. EEG originated with the observation of electrical activity in the brains of animals, first reported by a British physician, Richart Caton [2], in the late 19th century. Based on Caton's discovery, Hans Berger, a German psychiatrist, recorded the first human EEG from the human scalp with his ordinary radio equipment in 1924 [3]. It marked the beginning of EEG as a tool for studying the electrical activity of the human brain. In 1929, Berger published a paper on the topic, describing alpha and beta waves [4]. Over the next several decades, EEG became an increasingly important tool for understanding the brain and its functions. Advances in EEG technology allowed researchers to study the brain's electrical activity in more detail, leading to the discovery of various EEG patterns that were associated with different brain states, such as sleep, wakefulness, and epilepsy.

### 2.1.1 EEG Measurement

An electroencephalogram (EEG) is a graphical representation generated through the amplification and recording of spontaneous biopotentials in the cerebral cortex using an advanced instrument [3]. Nevertheless, the techniques for recording EEG are not restricted solely to the utilization of electrodes placed on the scalp. As an illustration, some research investigations have conducted EEG measurements from the ear [5, 6]. The depiction of this electrical activity is a planar graph of voltage versus time, with voltage serving as the vertical axis and time serving as the horizontal axis. The EEG is comprised of three fundamental components: the frequency (or period) of the brain waves, their amplitude, and their phase.

The electrodes used in EEG acquisition can be broadly categorized into three main groups, namely wet electrodes, semi-dry electrodes, and dry electrodes. Diverse attributes are associated with each category of measurement, rendering them appropriate for different research contexts and studies.

- **Wet electrodes** have low impedance, good reproducibility and stability, are most widely used in the clinical and research community, and are still the current gold standard.
- **Dry electrodes** have high impedance and poor stability, and the current research mainly focuses on solving such problems [7].
- **Semi-dry electrodes** are the current research hot-spot [8] due to their more comprehensive performance and convenient use.

To obtain accurate EEG data, steps must be taken to locate the position that generates the bioelectric signals, add more channels, and increase the sampling rate. This is because the EEG records the firing activity of some of the neurons involved in the activity, not the bioelectric signals of all the neurons involved in the activity.

The first human EEG was initially recorded by Berger using two electrodes applied to the scalp, one at the anterior and one at the posterior region of the skull [9]. Later, additional researchers brought attention to the fact that EEG activity varied greatly depending on the region of the scalp where it was recorded. The use of many electrodes and additional recording channels was prompted by the observation of various regional brain rhythms, but standardization of the recording techniques quickly became required in order to make the data produced comparable [9]. The first standardized system was the 10-20 system, published by Jasper in 1958 [10]. The 10-20 system consisted of 19 recording electrodes and 2 reference electrodes, as shown in Figure 2.1. It is ideal for the reference electrodes to have zero potential, meaning they should not have any bioelectric activity. The left and right earlobes are now frequently used as reference electrodes because, in actuality, there is hardly any zero potential on the surface of the human body. As a result, we can only select the areas of the body that move less and are less affected by different bioelectric fields. In addition, "10" and "20" in the system indicate that the actual separation between adjacent electrodes is either 10$\%$ or 20$\%$ of the overall separation between the left and right or front and rear of the skull. Regarding the names of the electrodes, the prefixes F, Fp, T, C, O, and P indicate the frontal polar, frontal, temporal, central, occipital, and parietal, respectively. The letter suffix "z" designates an electrode positioned along the midline, whereas an odd number signifies an electrode situated in the left hemisphere and an even number signifies an electrode positioned in the right hemisphere of the brain [11]. Fz and Cz are commonly used as the ground or common reference points for electrodes, and A1 and A2 are used as the opposite reference point [12].

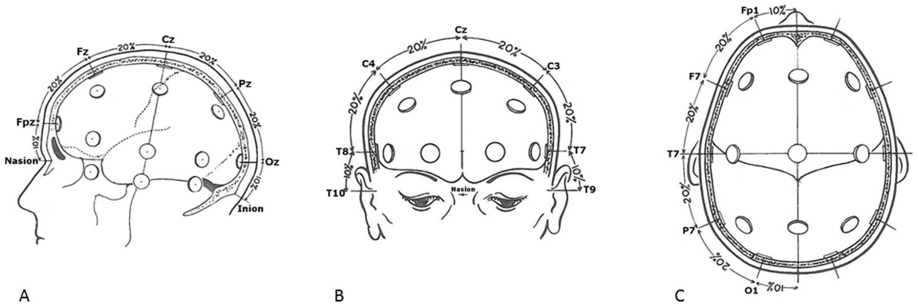


***Figure 2.1.*** *Standard electrode placement for the 10–20 system [12]*

The 10-10 system was introduced in the 1980s as an improvement to the 10-20 system, increasing the number of electrodes from 21 to 74 [11]. The left side in Figure 2.2 shows 21 electrode positions for the 10-20 system, while the right side shows 74 electrode positions for the 10-10 system. A comparison from Figure 2.2 shows that the blue electrodes in the right part are the 21 electrodes in the conventional 10-20 leads. Therefore, the 10-10 system can be described as an extension of the conventional 10-20 system, and the 10-10 system adds additional electrode positions to improve spatial resolution. The 10-10 system includes placement at 10% intervals between anatomical landmarks, resulting in more electrodes and finer scalp coverage [13].

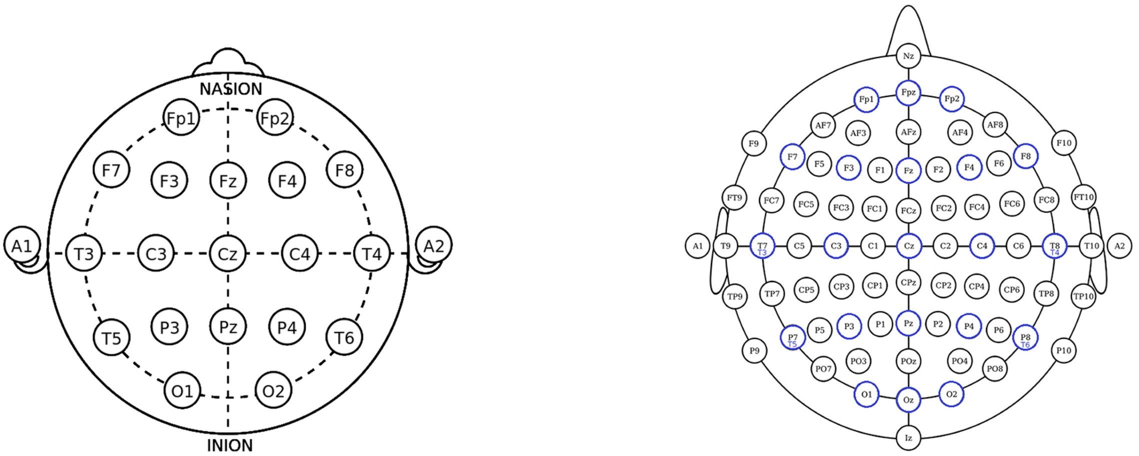


***Figure 2.2.*** *The left is the 10-20 International system of EEG electrode placement, and the right is the 10-10 International system of EEG electrode placement. [11]*

Over time, the 10-20 system attained global acknowledgment and established itself as the standard for the placement of EEG electrodes. While the 10-20 system and the 10-10 system provide general guidelines, individualized methods of electrode placement have also been developed. Personalized placement takes into account variations in head shape and anatomical landmarks to improve the accuracy and precision of EEG recordings [14].

### 2.1.2 Interpretation of the brain waves

Brain waves are generated by the electrical activities that emanate from the neural tissue within the brain. The transmission of signals by nerve cells generates brain electrical signals, commonly known as brain waves. The manifestation of brain waves displays distinct patterns and is correlated to some degree with the level of cerebral awareness. The variability of brain wave frequency is observed across various states, including but not limited to coma, excitement, and stress, and is distributed within the range of 0.5 Hz to 40 Hz. Brain waves are classified into distinct categories according to their corresponding frequencies, which comprise $\alpha$, $\beta$, $\delta$, $\theta$, and $\gamma$ waves. The aforementioned waves are correlated with distinct levels of awareness, cognitive functions, and psychological conditions, see Table 2.1.

At any given time, brain wave activity is not restricted to a singular type. Various brain rhythms can concurrently exist and interrelate with one another, with their respective magnitudes and configurations fluctuating according to an individual's cognitive state, engagement, and other pertinent aspects. Research on brainwaves offers valuable insights into the functioning of the brain, cognitive processes, and mental states.

***Table 2.1.*** *Summary of Brain Waves and Associated Mental States [15, 16, 17]*

| Brain Wave | Frequency Range | Interpretation |
|---|---|---|
| $\delta$ | 0.5-4 Hz | Deep sleep, physical healing, unconsciousness |
| $\theta$ | 4-8 Hz | Deep relaxation, daydreaming, creativity |
| $\alpha$ | 8-12 Hz | Relaxation, calmness, reflective states |
| $\beta$ | 12-40 Hz | Alertness, concentration, active thinking |
| $\gamma$ | 40-100 Hz | High-level cognitive processes, attention |

### 2.1.3 Artifacts

Artifacts are the noise recorded by the system that interferes with EEG data [18]. Prior to starting the collection and analysis of EEG data, it is critical to ensure that the data is as free from artifacts as possible, which means that the collected data should accurately represent brain activity. Consequently, it is critical to minimize and remove these artifacts to the greatest extent feasible.

### Source of Artifacts

The ability to identify artifacts is the first step in removing them. EEG artifacts can be classified according to their source, which may be physiological or external to the body (non-physiological). The most common ones are [18, 19, 20, 21]:

1. **Physiological artifacts**
    - Eye activity: Electrical potentials generated by eye movements, including saccades.
    - Muscle activity: Electrical activity from surrounding muscles.
    - Cardiac activity: Electrical activity associated with the heartbeat.
    - Sweat: Electrical changes due to moisture or sweat on the scalp.
    - Breathing: Electrical changes related to respiratory activity.
2. **Non-physiological/technical artifacts**
    - Electrode rejection: Displacement or detachment of electrodes from the scalp.
    - Cable movement: Artifacts caused by movement or tension in electrode cables.
    - Incorrect reference placement: Improper placement of the reference electrode.
    - AC and electromagnetic interference: Electrical noise from power lines or electromagnetic fields.
    - Body movements: Movement artifacts caused by the individual's body movements.

### Detection and Removal of EEG Artifacts

Accurate removal of artifacts in EEG involves a comprehensive approach that spans both the data collection and data analysis stages. By addressing artifacts at each step, researchers can minimize their impact and obtain high-quality EEG data for analysis, as shown below [22, 23, 24].

1. **Data collection phase:**
    - Proper preparation: Implement careful electrode placement, impedance checking, cable management, grounding, and noise reduction techniques, as discussed earlier.
    - Participant instructions: Provide clear instructions to individuals to minimize movements, avoid excessive eye blinks, and remain as still and relaxed as possible during the recording.

- High-quality recording equipment: Use reliable and properly calibrated EEG systems with adequate sampling rates and dynamic range to capture the EEG signals accurately.

2. **Data analysis phase:**

   - Rejection: The entirely automated statistical threshold method for EEG artifact rejection, as proposed by Nolan et al., is one example of the selection and rejection of EEG cycles containing artifacts. [25]
   - Filtering: Remove artifacts while maintaining as much EEG information as possible. For instance, simple linear filters, regression methods, adaptive filters with reference signals, the Wiener filter, and Bayesian filters.
   - Blind Source Separation (BSS): Decomposition of the EEG into linear combinations of signal sources based on different mathematical considerations. The most popular and useful techniques today are Independent Component Analysis (ICA) and Principle Component Analysis (PCA). In addition, Canonical Correlation Analysis (CCA) and EEG source imaging (ESI) are also commonly used BSS techniques.
   - Source Decomposition Methods: Each individual channel is decomposed into basic waveforms, and waveforms containing artifacts are removed to reconstruct a clean channel of the EEG signal. The main examples of these methods are Wavelet Transform (WT), and some less studied variants such as Empirical Mode Decomposition (EMD) or Nonlinear Node Decomposition (NND).

## 2.2 Cognitive load

### 2.2.1 Definition and theory

Cognitive load, which is also referred to as mental load, and brain load, pertains to the pace at which mental resources are expended in a work environment. Despite its significance, a precise and universally recognized definition of this concept remains elusive. According to the prevalent perspective, cognitive workload is a complex construct that encompasses multiple dimensions [26]. The 1977 NATO Human Factors Special Committee conference on "Mental workload: its theory and measurement" [27] posited that cognitive workload is contingent upon task demands and that the level of cognitive workload is determined by the workload itself. Cognitive load pertains to a multitude of factors, including task demands, temporal constraints, the operator’s cognitive capacity, exertion, performance, and various other elements. Cognitive load has been defined in various manners by scholars [27]. Wickens defines cognitive workload as "the relation between the (quantitative) demand for resources imposed by a task and the ability to supply those resources by the operator" [28]. Cain defines it as "a mental construct that reflects the mental strain that arise when performing tasks under specific environmental and operational conditions, coupled with the ability of the operator to respond to these demands" [29]. Abbass et al. differentiated various concepts in the literature [30], including mental load attributed to the work environment and mental load resulting from external environmental factors, such as the personal life of the operator. Two equations were utilized [30]:

$$CognitiveLoad \approx WorkLoad + EnvironmentalLoad \tag{2.1}$$

$$WorkLoad \approx TaskLoad + InterfaceLoad + OtherWorkRelatedFactors \tag{2.2}$$

Cognitive Load Theory (CLT) was initially introduced in 1988 by John Sweller [31], a cognitive psychologist affiliated with the University of New South Wales in Australia. CLT has been extensively researched by scholars worldwide since its inception [32]. CLT categorizes cognitive load into three different types: internal cognitive load, external cognitive load, and associated cognitive load. The three distinct forms of cognitive load are concurrently imposed on one another. According to CLT, the cognitive architecture of humans comprises two main components, namely working memory and long-term memory. Ericsson and Kintsch et al. [33] first introduced the concept of long-term memory in 1995.

Moreover, analyzing cognitive load and creating a theory of cognitive load are based on theories of cognitive resources. The most prevalent model of cognitive resources is Wickens' Multiple Resource Theory (MRT) model [28], which contends that there are really multiple resources available for processing information simultaneously rather than just one.

### 2.2.2 Methods to assess cognitive load

Currently, the assessment of cognitive workload can be categorized into two distinct types: subjective and objective measures [34]:

- Subjective assessment refers primarily to user self-assessment and involves technical instruments such as surveys and questionnaires. Commonly used questionnaires include the National Aeronautics and Space Administration-Task Load Index (NASA-TLX) [35] and Subjective Workload Assessment Technique (SWAT) [36].
- Objective assessments can be carried out from a neurophysiological, physiological, and behavioral perspective. EEG and functional near-infrared spectroscopy (fNIRS) are the major neurophysiological methods used to assess cognitive load [37]. Electrocardiography (ECG), respiration, electrodermal activity (EDA), and eye movement measurements are examples of physiological parameters. Behavioral metrics include keyboard dynamics, mouse tracking, and body positioning [38], as well as accuracy, response time, and speed of completion in relation to real task performance.

# 3. TRANSFORMER NETWORK: BASIC ARCHITECTURE AND OPERATION

In recent years, the transformer network architecture has become increasingly prevalent in the field of deep learning. Originally published by Vaswani et al. in 2017 [39], transformers have shown to be extremely effective in a range of natural language processing (NLP) applications, including text classification and language translation. Furthermore, The transformer network is the first transformation model that excludes the use of sequence-aligned recurrent or convolutional neural networks. Instead, it computes the representation of inputs and outputs only through self-attention [39]. Transformer has not only revolutionized the field of NLP but also shown remarkable success in other areas of machine learning.

This chapter will discuss the architectures and operations of transformer networks in NLP, with a focus on their unique features that set them apart from other neural network architectures.

## 3.1 Transformer Network Architecture

The transformer network is constructed using an encoder-decoder architecture, which involves the stacking of multiple identical encoders and decoders. The encoders process the input sequence and the decodes generate the output sequence. In a typical sequence-to-sequence converter model used for tasks such as machine translation, the number of cells in both decoder and encoder stacks is the same, as shown in Figure 3.1. This is because capturing and transmitting correlation information between the input and output sequences is facilitated by having an equivalent number of layers in the encoder and decoder. The following describes the workflow of the transformer network:

1. The first step is to obtain the representation vector $X$ for each word in the input phrase. The $X$ consists of the embedding of the word and the embedding of the word position added together.
2. The encoder receives the word representation vector, and after six encoder blocks, the information matrix for all the words in the phrase is acquired. The word vector is represented by the symbol $X_{n\times d}$, where $n$ represents the number of words in the

phrase and $d$ represents the dimension of the vector. The output matrix from each encoder block is of the same dimension as the input matrix.

3. The next word is translated by the decoder based on the word that is presently being translated using the encoded information matrix output by the encoder.

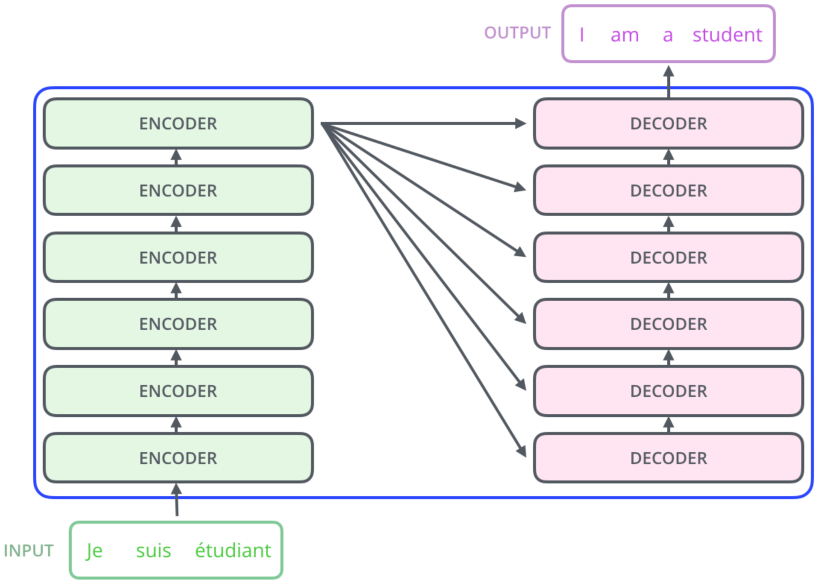


***Figure 3.1.*** *The transformer encoder-decoder stack [40]*

The preceding part illustrates the architecture of the transformer and the overall sequence of its application. Subsequent sections will expand on the intricacies of the individual components therein.

## 3.2 Transformer Network Operations

### 3.2.1 Input to Transformer

Summing the initial input embedding and the positional encoding yields the input to the encoder section in Transformer, as shown in Figure 3.2 below.

- **Embedding:** Each symbol in the input sequence is first transformed by the transformer into a fixed-size vector representation. Embedding can be obtained in a number of methods, including pre-training with algorithms such as Word2Vec and Glove [41], or training in Transformer.
- **Positional Encoding:** The transformer employs positional encoding in addition to word embedding to describe the placement of words in a phrase. Since the transformer uses global information, it is unable to use the words' sequence information, which is crucial for NLP. As a result, the transformer maintains the absolute or relative positions of the words within the series by using positional encoding. Moreover, the dimension of the positional encoding and the embedding should be the same,

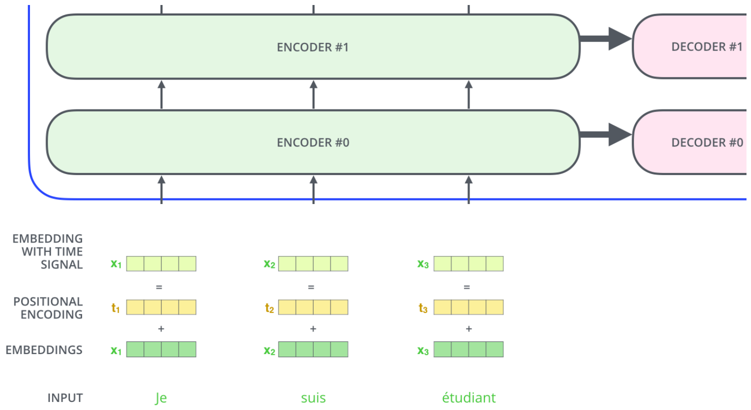


***Figure 3.2.*** *Positional Encoding [40]*

so that they can be added to each other [39]. Positional encoding can be acquired through training or by applying an algorithm. Transformer uses the latter, which is computed as follows [39]:

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d}) \tag{3.1}$$

$$PE_{(pos,2i+2)} = cos(pos/1000^{2i/d}) \tag{3.2}$$

where $pos$ is the position of the word in the sentence, $d$ is the dimension of positional encoding, $2i$ indicates the even dimension and $2i+1$ represents the odd dimension.

Even for sequences with varying lengths, the unique positional encoding of each location in the sequence is guaranteed by the employment of sine and cosine functions with distinct frequencies. The positional encoding is subsequently appended to the input embeddings prior to their input into the encoder of the transformer.

### 3.2.2 Attention

Our ability to form snap judgments about what we see is due to the fact that our brains automatically zero down on the most salient features of an item rather than forcing us to examine it in its entirety. It is on this theory that the attention mechanism was developed. With the encoder block on the left and the decoder block on the right, the accompanying Figure 3.3 illustrates the internal construction of the transformer in the [39]. One multi-head attention is present in the Encoder block, whereas two multi-head attentions are present in the Decoder block, one of which is masked. Since the key component of the transformer network architecture is self-attention, this section will mainly discuss self-attention and multi-head attention.

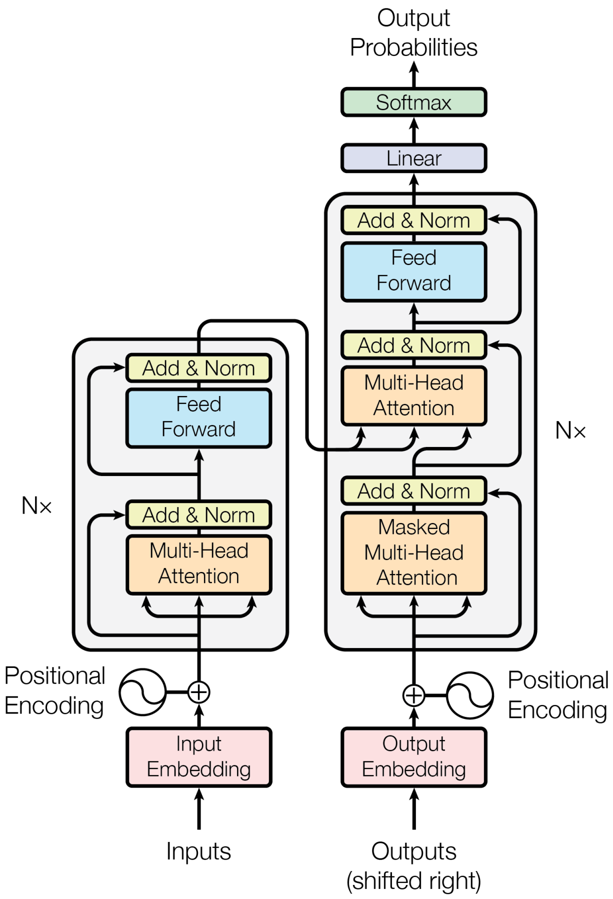


***Figure 3.3.*** *The Transformer - model architecture [39]*

### Self-Attention

Figure 3.4 illustrates the architecture of self-attention, which utilizes the matrices $Q$ (query), $K$ (key), and $V$ (value) during computation. In practical applications, self-attention is typically provided with either the input subsequent to embedding and positional encoding or the output originating from the preceding encoder block. The values of $Q$, $K$, and $V$ are obtained through a linear conversion of the input of self-attention.

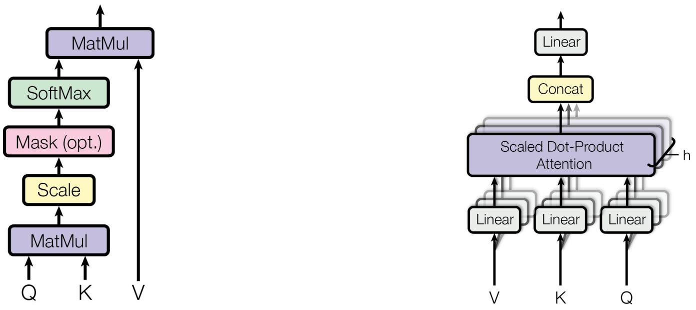


***Figure 3.4.*** *The left is scaled dot-product attention, and the right is multi-head attention [39]*

As shown on the left in Figure 3.5, the input matrix $X$, after embedding and positional encoding, is multiplied by the trained weight matrices $W^Q$, $W^K$, $W^V$. Afterwards, the attention calculation is performed based on the following formula [39]:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3.3}$$

where $\sqrt{d_k}$ is the number of columns of the Q, K matrix, which is the dimension of the vector.

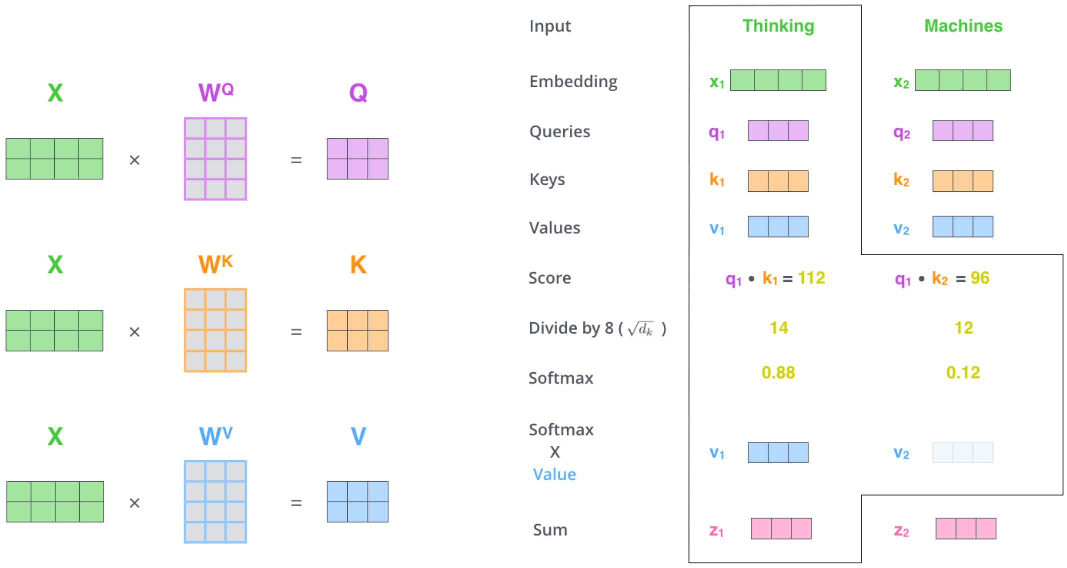


***Figure 3.5.*** *Self-Attention Calculation [40]*

After the $Q$, $K$, and $V$ matrices are determined, the score is derived through the multiplication of $Q$ and $K$ as shown on the right-hand side of Figure 3.5, which are capable of representing the degree of attention between words. The softmax function is employed to determine the attention coefficient of each word relative to the other words subsequent to the acquisition of $QK^T$. Upon obtaining the softmax matrix, it can be subjected to multiplication with $V$, resulting in the ultimate output $Z$. Ultimately, a $Z$-matrix is derived, serving as the resultant product of the self-attention layer.

### Multi-Head Attention

The multi-head attention is comprised of multiple self-attention components, as depicted in the right side of Figure 3.5 presented in the [39]. Following the embedding and positional encoding process, the input matrix denoted as $X$ is passed to each of the $h$ different self-attention. The term "different self-attention" refers to $h$ groups of $Q$, $K$, $V$ obtained by multiplying the input matrix $X$ with the $h$ groups of $W^Q$, $W^K$, $W^V$ weight matrices. Each group $Q$, $K$, $V$ will eventually result in a weight matrix $Z$. The final output $Z$ of multi-head attention is created by concatenating $h$ weight matrices $Z$ and passing them into a Linear layer.

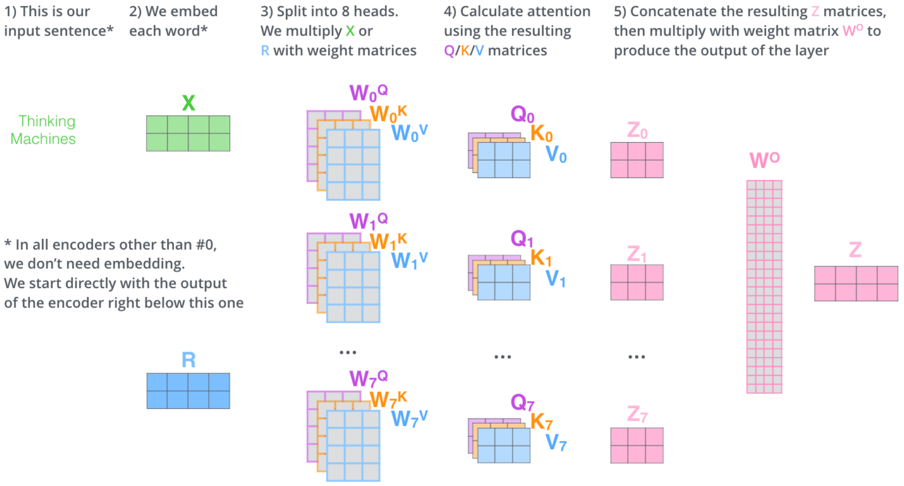

1) This is our input sentence*
2) We embed each word*
3) Split into 8 heads. We multiply X or R with weight matrices
4) Calculate attention using the resulting Q/K/V matrices
5) Concatenate the resulting Z matrices, then multiply with weight matrix $W^O$ to produce the output of the layer
Thinking Machines
X
* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one
R
$W_0^Q$
$W_0^K$
$W_0^V$
$W_1^Q$
$W_1^K$
$W_1^V$
...
$W_7^Q$
$W_7^K$
$W_7^V$
$Q_0$
$K_0$
$V_0$
$Q_1$
$K_1$
$V_1$
...
$Q_7$
$K_7$
$V_7$
$Z_0$
$Z_1$
...
$Z_7$
$W^O$
Z

***Figure 3.6.*** *Multi-Headed Self-Attention [40]*

### 3.2.3 Encoder

The left-hand side of Figure 3.3 represents the encoder blocks, each of which consists of a multi-head attention layer, 2 add&norm layers, and a feed-forward layer. In the preceding section, the calculation process of multi-head attention has been comprehended. The sections that follow will be explained: add&norm and feed-forward.

#### Add&Norm

The add&norm layer is comprised of two distinct components, namely add and norm, which are computed in the following formulas:

$$LayerNorm(X + MultiHeadAttention(X)) \tag{3.4}$$

$$LayerNorm(X + FeedForward(X)) \tag{3.5}$$

where $X$ is the input to multi-head attention or feed-forward, while $MultiHeadAttention(X)$ and $FeedForward(X)$ are the outputs of multi-head attention or feed-forward.

The term "Add" pertains to the operation of combining $X$ and $MultiHeadAttention(X)$, along with a residual connection. Conversely, "Norm" refers to the process of Layer Normalization, which standardizes the inputs of every layer of neurons to possess identical mean-variance values, thereby expediting the convergence process [42].

### Feed Forward

The feed-forward layer comprises two fully connected layers. The first layer employs ReLu as its activation function, while the second layer does not have any activation function. This can be represented mathematically as follows [39]:

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b2 \tag{3.6}$$

where $X$ is the input and feed-forward ends up with an output matrix that has the same dimensions as $X$. The ReLU function is a segmented linear function that treats negative values as 0.

### Encoder Operation

The multi-head attention, feed-forward, and add&norm operations that were explained before are used to form an encoder block. This block takes in a matrix $X_{(n*d)}$ as its input and produces a matrix $O_{(n*d)}$ as its output. Multiple encoder blocks can be stacked together to produce an encoder. A vector containing the words is the first input to the first encoder block. The output of the previous encoder block is used as the input for the next encoder blocks. The final encoder block produces the output matrix for the entire encoder, which is subsequently utilized in the decoder.

## 3.2.4 Decoder

The section on the right-hand side of Figure 3.3 depicts the decoder component of the transformer, which has a resemblance to the encoder module, but with particular distinctions:

- There are two multi-head attention layers present.
- The first multi-head attention layer uses a masked operation.
- The output matrix of the encoder is used to construct the $K$ and $V$ matrices of the second multi-head attention layer, whereas the output of the previous decoder block is used to calculate $Q$.
- For each translated word, the probability is computed by a softmax layer.

### The First Multi-Head Attention

Due to the sequential nature of the translation process, the initial multi-head attention of the decoder block makes use of the masked operation. This is the case because it is necessary to translate the $i_{th}$ word before translating the $(i+1)_{th}$ word. Through the use of the masked operation, the $i_{th}$ word is shielded from the information that comes after the $(i+1)_{th}$ word.

The process of executing the masking operation entails the utilization of the masking matrix to mask the information subsequent to each word, prior to the softmax of self-attention. This results in the generation of a masking $QK^T$. Upon conducting the softmax operation on the given basis, the attention score of the $i_{th}$ word towards the other words is 0. The subsequent procedure is identical to the preceding self-attention. Ultimately, the masking self-attention produces an output matrix denoted as $Z_i$, which is subsequently concatenated with multiple outputs through multi-head attention, similar to the encoder. The resultant matrix $Z$ from "the first multi-head attention" operation is calculated to possess identical dimensions as the input matrix $X$.

### The Second Multi-Head Attention

The main difference between the first multi-head attention and the second multi-head attention lies in the fact that the $K$ and $V$ matrixes for self-attention are derived not from the output of the preceding decoder block, but rather from the encoder's output matrix. The values of $K$ and $V$ are derived from the output matrix of the encoder, while $Q$ is obtained from the output $Z$ of the preceding decoder block, or from the input matrix $X$ if it is the first decoder block. The subsequent computations follow the previously outlined methodology. One benefit of this approach is that during the decoding phase, individual words have access to the collective information from all the words in the encoding phase without requiring any masking.

### Softmax: predicting output words

The last step of the decoder block requires using softmax to anticipate the subsequent word. In the prior network layer, the final result $Z$ can be obtained because the presence of masking ensures that the output $Z_0$ of the $i_{th}$ word contains only the information related to the $i_{th}$ word. The softmax function is utilized to predict the subsequent word by taking into account each row of the output matrix.

## 3.3 Discussion of Transformer Network Applications

Despite its current popularity, the Transformer model is not without notable imperfections and limitations. In NLP, tasks that involve lengthy inputs, such as those at the chapter level, often pose a significant computational challenge for the transformer model. This is due to the excessive length of the input, which can result in a considerable decrease in processing speed. Zihang Dai et al. introduced the Transformer-XL architecture in 2019 [43], which effectively addresses the challenge of processing lengthy input sequences. The transformer model has demonstrated efficacy not only within the domain of NLP but has also garnered widespread adoption across various other disciplines. Therefore, the transformer model needs to be adapted differently to cope with different challenges.

Since the birth of the transformer model, subsequent research and development has evolved in three main areas: model efficiency, model generalization, and model adaptability [44].

- **Model efficiency** focuses on reducing the computational and memory complexity by adjusting the content of the transformer structural module and the overall structure. For example, improvement of the multi-headed attention mechanism, adjustment of the layer normalization approach, and transformer lightweight.
- **Model generalization** is mainly the introduction of structural bias or regularization, pre-training of large-scale unlabeled data, and so on. Transfomer-based pre-training models can be divided into three main categories: those that use only encoders, those that use only decoders, and those that use both.
- **Model adaptability** is to apply the transformer model to more fields. Transformer model is first used in the field of NLP, such as machine translation and the subsequent Bidirectional Encoder Representations from Transformer (BERT) and Generative Pre-trained Transformer (GPT) series. Transformer is also used in the field of computer vision for image classification, object detection, image generation, and video processing tasks, such as Vision Transformer (ViT). Furthermore, Transformer has also been applied to the field of speech for tasks such as speech recognition, speech synthesis, speech enhancement, and music generation. Multimodal scenarios consisting of NLP, vision, and speech are also hot directions for Transformer applications in recent years, such as visual question and answer, visual common sense reasoning, speech-to-text translation, and text-to-image generation. In addition to the usual AI scenarios of NLP, vision, and speech, Transformer has also been applied to the field of psychology.

This chapter thoroughly overviews the transformer's structure and functionality, delving into its primary constituents, including self-attention mechanisms and position encoding, which endow the transformer with exceptional aptitudes across diverse domains. The subsequent chapter will underscore the significance of the transformer in EEG research and its capacity to enhance our comprehension of brain dynamics. This will be achieved through a comprehensive review of literature pertaining to the utilization of the transformer in EEG signal analysis.

# 4. TRANSFORMER NETWORK IN ANALYZING EEG SIGNAL: LITERATURE REVIEW

EEG data analysis is essential for comprehending the intricate dynamics of the human brain. Many machine learning and signal processing methods have been employed over time to glean important information from EEG data. The transformer model, which was first used in the study of NLP, has lately drawn a lot of interest for its potential to improve EEG analysis. Its ability to capture complex patterns and dependencies in sequences has led to early applications in a variety of fields such as emotion detection, sleep quality assessment, and cognitive load.

This chapter explores four methods for applying transformer networks to EEG analysis: utilizing them alone, in conjunction with other deep learning networks including capsule networks and convolutional neural networks (CNN). In the direction of EEG analysis, the network architecture that combines transformer and CNN is more prevalent. Consequently, the following includes two distinct combinations of transformer networks and CNN with their respective applications.

## 4.1 Study 1

For the classification of the raw EEG data, Siddhad et al. first proposed the application of transformer networks. [45]. They assessed its performance in comparison to established deep learning networks. The experimentation involved the utilization of two distinct datasets: one sourced locally, focusing on age and gender, and the other being the open-access mental workload dataset named the Simultaneous Task EEG Workload (STEW) Dataset [46]. Notably, the proposed framework got top-notch accuracy on both datasets. This shows how transformer networks can be used to learn the features and improve the accuracy of EEG data classification.

### Data Pre-processing

In this study, the researchers utilized two distinct datasets for their investigations. The first dataset was locally collected and focused on demographic characteristics such as age and gender. This dataset comprised resting-state EEG data obtained from a sample of 60 individuals, evenly divided between 30 males and 30 females. The EEG data was recorded using a 14-channel EEG system operating at a sampling rate of 128 Hz. The second dataset employed in this study was the open-access mental workload dataset known as STEW [46]. The dataset consisted of unprocessed EEG data obtained from 48 participants who engaged in a simultaneous capacity (SIMKAP) multitasking test-based multitasking workload experiment [47]. Two different experiment types are included in the dataset: "No task" and "SIMKAP-based multitasking activity." The "No task" experiment has been divided into low and high workload levels, whereas the "SIMKAP-based multitasking activity" has been divided into three workload levels, namely low, moderate, and high, based on the subjects' ratings [48]. The EEG signals were captured using the Emotiv EPOC EEG headset, equipped with 14 channels (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4) with 2 reference channels (CMS, DRL), and sampled at a rate of 128 Hz [46].

To prepare the EEG data for analysis, a series of preprocessing steps were undertaken. These steps included bandpass filtering, epoch segmentation to isolate relevant data segments, the removal of segments with poor quality, and the application of ICA to mitigate artifacts. Subsequently, the preprocessed EEG data was divided into fixed-length segments, serving as input for the transformer network. Each segment of EEG data was treated as a chronological sequence of data samples, with each sample representing the voltage measurement from a specific electrode at a given time point. Through the implementation of embedding and positional encoding techniques, this sequence of data samples was transformed into a sequence of feature vectors. The resulting sequence of feature vectors was then input into the transformer network, a critical component responsible for encoding the input sequence into a sequence of uniform feature vectors of fixed dimensions.

### Network Architecture

To better adapt the original transformer network architecture for EEG data categorization tasks, the authors made some modifications. When dealing with tasks involving natural language processing and an input consisting of a string of words or tokens, the initial transformer network was created. On the other hand, the network receives a series of data samples as input, whereas EEG data are continuous time-series signals. In order to properly handle this kind of input, the authors had to change the transformer network's architecture. To be more precise, they employed an embedding layer to convert the input sequence of data samples into a sequence of feature vectors, and a position encoding layer was then utilized to add information about each feature vector's position within the input sequence. For the STEW dataset and the age and gender dataset, they employed a stack of four transformer encoder layers in their suggested architecture. A positionally fully connected feed-forward network and a multi-head self-attention mechanism make up each encoder layer's two sub-layers. An essential component of classifying EEG data is the network's ability to recognize interdependencies among the segments of the input sequence, which is made possible by the multi-head self-attention mechanism. In order to handle the self-attention mechanism's output, a feed-forward neural network was then added. This neural network aids in transforming the network to capture the intricate nonlinear correlations between the input variables, which are explained in the preceding chapter.

The researchers employed two distinct transformer networks to process the two datasets under investigation. For Age and Gender classification, the experiments were conducted separately for the age and gender datasets. The network architecture employed for gender classification (2 classes) is visually depicted on the left of the accompanying Figure 4.1. For the age classification task (6 classes), the number of nodes in the last layer of the model, the Dense layer, also known as the fully connected layer, was modified to 6. Furthermore, the attention heads were augmented to a total of 8 to enhance the network's capabilities. Likewise, with regards to the STEW dataset, the research endeavors extended to encompass examinations of the "No task" (2 classes) and the "SIMKAP-based multitasking activity" (3 classes) datasets. The architectural configuration deployed for the "No task" is visually presented on the right-hand side of the accompanying Figure 4.1. Drawing from this framework, the final layer was strategically modified by adjusting the number of nodes in the fully connected layer to 3. This modification was carried out to align the architecture with the complexity of SIMKAP multitasking classification.

### Result

The outcomes of the experiments on the age and gender dataset highlight the effectiveness of the proposed transformer network. It achieved a remarkable accuracy of 94.68$\%$ for gender classification and 87.63$\%$ for age classification. Comparisons were made with leading methods by Zhu et al. [49] and Zheng and Chen [50], demonstrating that the proposed transformer network outperforms these benchmarks in both gender and age classification tasks.

Regarding the STEW dataset experiments, the proposed transformer network yielded impressive results. It achieved 95.32$\%$ accuracy for the "No task" classification and 89.01$\%$ for the "SIMKAP multi-task" classification. Comparisons with methods by Zhang et al. [49] further underscored the superiority of the proposed transformer network across both classification scenarios.

In general, the findings of the research indicate that transformer networks exhibit superior performance in the classification of EEG data, particularly in the absence of any requirement for manual feature extraction. Nevertheless, the authors highlight that the positional encoding employed in transformer networks designed for text processing is not specifically optimized for EEG data, and that features, which can be more effective in certain situations, are not utilized. The authors suggest that further research is needed to validate the proposed method with different datasets and more comparisons.

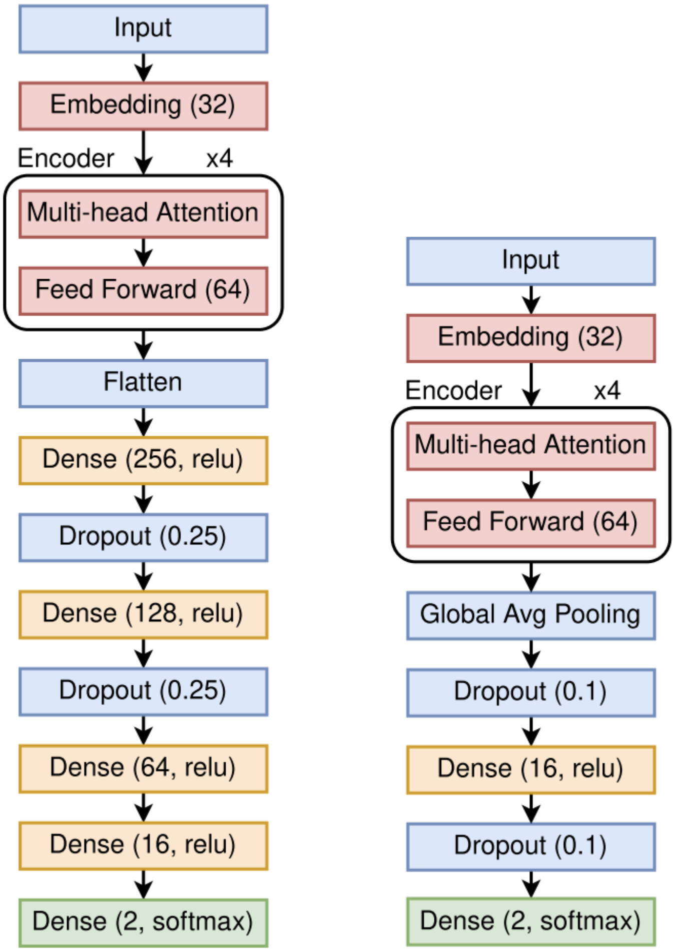


***Figure 4.1.*** *The architecture for Age and Gender and STEW dataset are on the left and right, respectively. [45]*

## 4.2 Study 2

Wei et al. proposed a deep learning model named Transformer Capsule Network (TC-Net) [51] for emotion recognition from EEG signals. It combines the power of transformers and capsule networks to capture local and global contextual information from EEG signals.

### Data Pre-processing

Before feeding the signal into TC-Net, they pre-processed the EEG signal by applying segmentation, baseline removal, and continuous wavelet transform (CWT) as shown in Figure 4.2.

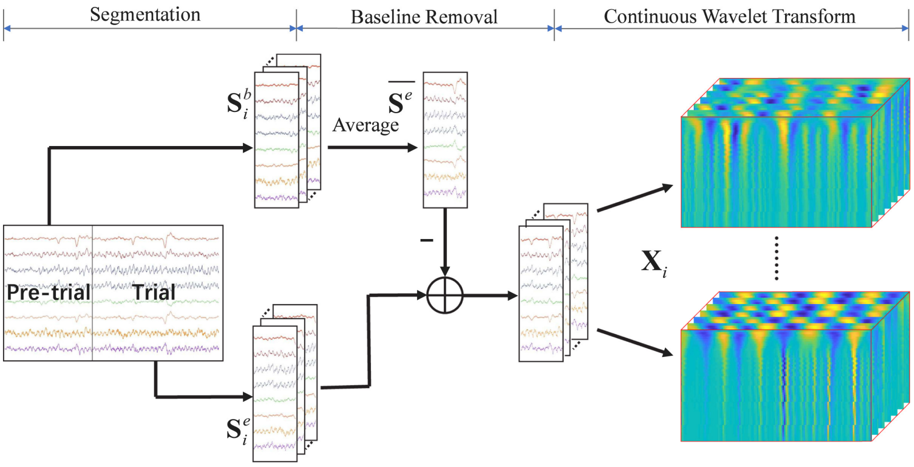


***Figure 4.2.*** *The flow chart of signal preprocessing in TC-Net [51]*

In the segmentation stage, a sliding window slices the EEG signals into 1-second intervals. The notation for each segment is $S_i$, where $i$ = 1, 2, ..., N, where N is the number of segments. Additionally, $S_i$ is an element of the set $\mathbb{R}^{C \times L}$, where $C$ represents the number of EEG channels and $L$ represents the window length. After segmentation, they applied baseline removal since many previous studies have employed the practice of removing baseline signals to improve performance outcomes. For each segment that underwent baseline removal, they performed CWT for each channel separately. CWT enables the wavelet to capture the innate frequency characteristics by sliding it over the signals along the time dimension. The CWT mechanism is responsible for this conversion of raw signals to the time–frequency domain.

### Network Architecture

As illustrated in Figure 4.3, the model comprises three primary modules subsequent to the data processing: the patch partitioning module, the EEG transformer module, and the emotion capsule module.

- **Patch partition module:** Pre-processed EEG signals are initially segmented into small patches using convolution operations, with patch size determined by frequency and temporal resolution considerations. During convolution, both the kernel size and stride are aligned with the patch size. As a result, EEG signals from each channel are partitioned into adjacent non-overlapping patches. In this manner, the patches sustain the original frequency and temporal attributes. These patches are then processed through an EEG transformer module consisting of various blocks.
- **EEG transformer module:** The feature extraction is accomplished by the EEG transformer module, which consists of four primary components: the window partition block, the transformer block, the window reverse block, and the patch merging block. The window partition block divides tokens into non-overlapping feature

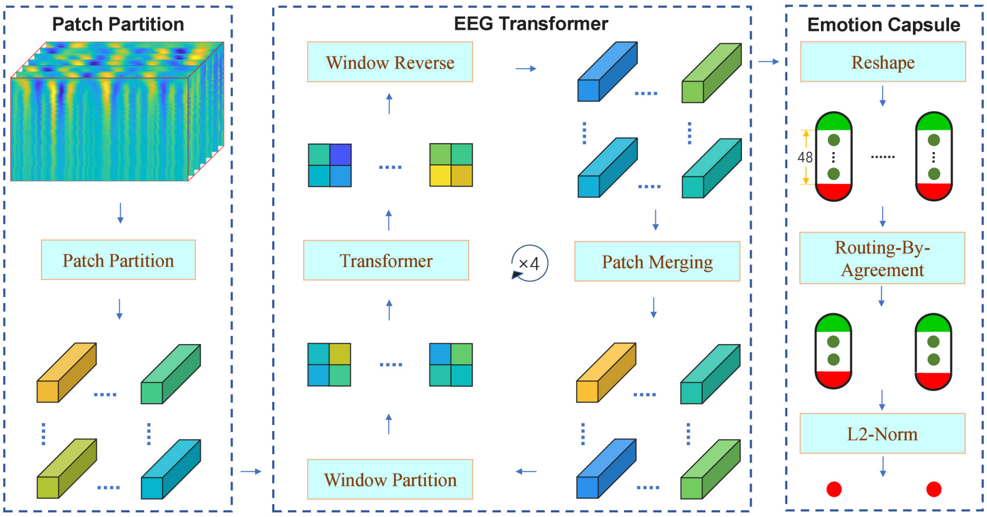


***Figure 4.3.*** *The architecture of Transformer Capsule Network (TC-Net) [51]*

windows, maintaining computational efficiency. A series of feature vectors are produced as a result of a subsequent transformer block employing multi-head self-attention to capture temporal dependencies. The window reverse block ensures the preservation of the original temporal order. In the patch merging block, neighboring patches are merged to capture local features, and the resulting patches are passed to the next module layer.

To preserve the inherent local attributes of EEG signals, the researchers introduce a novel patch merging strategy called EEG Patch Merging (EEG-PM), illustrated in Figure 4.4. This convolutional operation employs a larger kernel and increased output channels, doubling the feature maps while halving their resolution. In order to derive high-level emotional features, the EEG Transformer module undergoes four iterations. Prior to the patch merging block, the feature map is generated in the $4_{th}$ cycle. This comprehensive process facilitates the extraction of discriminative features from pre-processed EEG signals for classification.

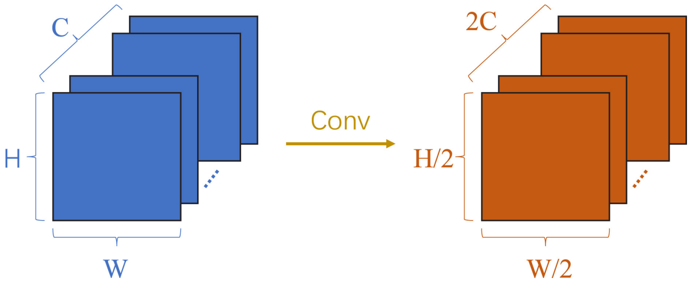


***Figure 4.4.*** *The EEG-PM strategy [51]*

- **Emotion capsule module:** The emotion capsule module plays a pivotal role. It is responsible for categorizing EEG features into two emotional states. To ensure a comprehensive representation of relationships between different channels within the feature map, the number of capsules is set to match the number of feature map channels. This alignment facilitates the subsequent processing steps. In this process, EEG features are transformed into capsules, with each capsule containing 8 neurons. The introduction of the dynamic routing-by-agreement mechanism aids in capturing the intricate relationships that exist among the various feature map channels. This mechanism enhances the model's ability to capture essential information embedded within the EEG data. The ultimate classification results are derived by evaluating the L2-Norm of each capsule after the dynamic routing-by-agreement process. This step provides a solid basis for producing accurate and reliable classification outcomes, effectively summarizing the emotional states inferred from the EEG features.

#### Result

Experiments were performed by the authors on two widely used datasets that pertain to emotion recognition from EEG signals: the database for emotion analysis using physiological signals (DEAP) [52] and the database for emotion recognition through EEG and ECG signals (DREAMER) [53]. EEG signals from 32 electrodes were recorded in DEAP in accordance with the international 10-20 system while subjects viewed 40 one-minute music videos. EEG signals from 14 electrodes were recorded in DREAMER in accordance with the standard 10-20 system while subjects viewed 18 movie excerpts.

A subject-dependent 10-fold cross-validation method was employed to assess the efficacy of TC-Net. In this method, both the training and testing datasets originated from the same subject. The results showed that TC-Net outperformed several cutting-edge approaches on both datasets in terms of accuracy. Specifically, on the DEAP dataset, TC-Net achieved an accuracy of 98.76$\%$ for valence, 98.81$\%$ for arousal, and 98.82$\%$ for dominance. On the DREAMER dataset, TC-Net achieved an accuracy of 98.59$\%$ for valence, 98.61$\%$ for arousal, and 98.67$\%$ for dominance. The outcomes of this study illustrate the efficacy of TC-Net in emotion recognition using EEG and its potential for practical implementations.

## 4.3 Study 3

Guo et al. present an innovative neural network model known as depthwise convolutional transformer (DCoT), which integrates depthwise convolution and Transformer encoders for EEG-based emotion recognition [54]. The researchers investigate the intricate relationship between emotion recognition and individual EEG channels, enhancing interpretability by visually presenting the extracted features. Impressively, the DCoT model achieves notable classification accuracy while also unveiling the importance of distinct EEG channels in representing emotional states through brain map visualization. The discourse within the article also contemplates the practical applications of this technology in real-world scenarios, particularly its potential to reduce equipment and computational costs.

### Data Pre-processing

To validate their approach, the authors utilized the SJTU Emotion EEG Dataset (SEED) [55]. This dataset contains EEG data collected from 15 participants, comprising 7 males and 8 females, with an average age of 23.27 ± 2.37 years. These participants took part in emotional experiments where they watched 15 emotionally evocative film clips, each about 4 minutes in length. These clips were carefully chosen to induce positive, neutral, and negative emotions, with five corresponding clips for each emotion. After watching each clip, participants were asked to provide emotional feedback through questionnaires. This experiment was repeated three times, every two weeks, resulting in a total of 45 trials per participant across three sessions.

For the SEED dataset, a systematic preprocessing approach was utilized. This included downsampling the raw EEG signals from 1000 Hz to 200 Hz, applying ICA to eliminate unwanted signals, implementing a bandpass filter between 0 and 50 Hz for noise reduction, segmenting EEG signals into 10-second intervals aligned with movie clips, selecting relevant data epochs within the range of 1000 to 37000, and extracting distinct frequency rhythms ($\delta$, $\theta$, $\alpha$, $\beta$, $\gamma$) from these epochs. This comprehensive preprocessing method establishes a strong foundation for insightful analyses of the SEED dataset.

Before inputting data into the model, the authors harnessed differential entropy (DE) as a nonlinear entropy metric for manual feature extraction. Differential entropy is particularly effective in recognizing EEG signals, especially those related to emotions [56]. The standard definition of differential entropy involves integrating the probability density function of a random variable. Given that EEG signals tend to follow a Gaussian distribution [56], their differential entropy can be determined as the logarithmic energy spectrum within a specific frequency range. This process involved the extraction of DE features across five EEG frequency bands ($\delta$, $\theta$, $\alpha$, $\beta$, $\gamma$), facilitated by a 256-point Short-Time Fourier Transform using a non-overlapping Hanning window of one second. As a result, DE features were derived with dimensions of $N \times 1 \times C$ per second, where $N$ represents the number of EEG channels and $C$ stands for the EEG bands. For a sample of $T$ seconds, the authors obtained DE features with dimensions of $N \times T \times C$.

### Network Architecture

The model design is founded upon the original Vision Transformer model, which was initially introduced in the domain of computer vision (CV). The DCoT model is made up of numerous components, including a depthwise convolution (DW-CONV) layer, position embeddings, learnable embeddings, transformer encoders, and linear layers, see Fig-

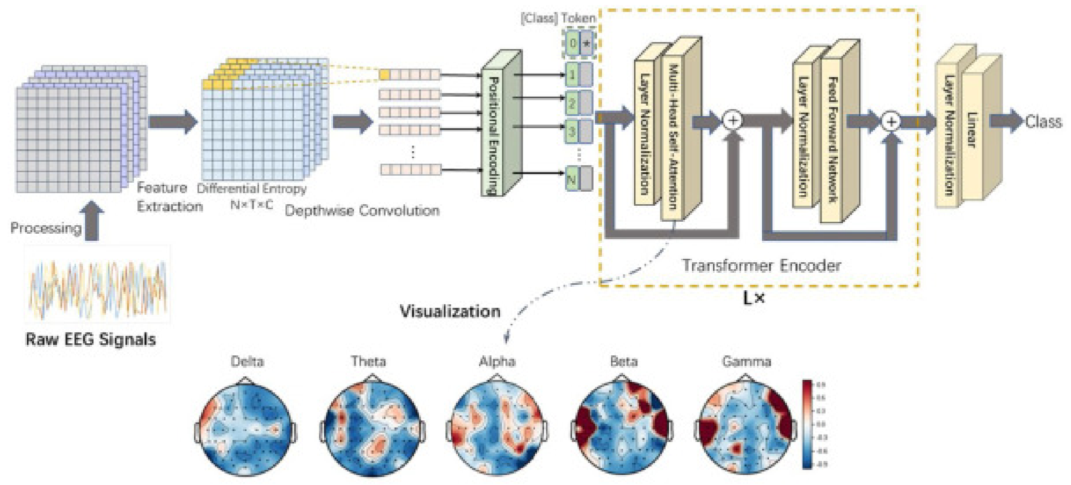


***Figure 4.5.*** *The architecture of the proposed DCoT model [54]*

ure 4.5.

- **DW-CONV:** The DCoT model leverages a DW-CONV layer for processing the input DE features, playing a pivotal role in the model. This layer facilitates the extraction of comprehensive information from multi-frequency data. While conventional Transformer models segment input data into patches and reshaped vectors, they can disrupt temporal coherence and lead to undesirable interference. To counteract this, the convolution layer is introduced to enhance coherence and fuse DE features across frequencies. It also aids in capturing local and frequency domain features, thus streamlining subsequent computations and enhancing emotion recognition results.

  In ensuring the independence of EEG channel features, a depthwise convolution layer is utilized due to its efficacy in extracting features from distinct color channels in CV. In this adaptation, EEG channels are treated as image channels. The depthwise convolution layer employs $N$ kernels with dimensions of $C \times f$, where $C$ represents the frequency domain, $f$ denotes the time domain, and the stride is $s$. The outcome is a feature matrix $X_0$ with dimensions $N \times D_f$, where $D_f$ is derived from the temporal length $T$ and stride s. To uphold channel independence, the EEG feature matrix $X_0$ is segmented into one-dimensional vectors based on EEG channels, resulting in $N$ inputs for the subsequent Transformer encoder.

- **Positional embedding and learnable embedding:** The authors introduced one-dimensional learnable position embeddings into the input sequences of the encoders to encode positional information. These position embeddings help maintain the sequence orders of different channel feature vectors. Simultaneously, an additional [class] token, denoted by a box with '$*$', is added to the input sequence, serving as the learnable embedding, as shown in Figure 4.5. This learnable embedding acts as the representation of EEG features. The output from the Transformer encoder represents the learnable embedding, effectively capturing the input feature representation.

- **Transformer encoder:** In the context of the DCoT model, the Transformer encoder holds substantial importance, serving as a key element to capture temporal dependencies among the extracted features. The Transformer encoder configuration comprises layer normalizations (LN), multi-head self-attention (MSA) layers, and feed-forward networks (FFN). The investigation into intrinsic relationships among EEG channels occurs predominantly through a sequence of stacked Transformer encoders. Each encoder is composed of two primary components: MSA and FFN. The structure entails residual connections encircling MSA and FFN, as depicted in Figure 4.5, subsequently followed by layer normalization. This study employs five encoder layers.

  The MSA layer is tasked with capturing the interplay between informative signals, while the FFN layer focuses on capturing the fundamental features within distinct EEG frequency bands. The attention mechanism holds the capacity to identify pivotal EEG channels for emotion recognition, contributing to the interpretability of DCoT's learning process. The output from the Transformer encoder is then directed through a linear layer to yield the final classification outcome.

#### Result

The experiment results using the SEED dataset affirm the strong performance of the proposed DCoT model in EEG-based emotion recognition. The authors conducted both subject-dependent and subject-independent evaluations, showcasing impressive average accuracies across different classification tasks. In subject-dependent experiments, accuracies reached remarkable levels of 99.82$\%$ for two tasks and 93.83$\%$ for three tasks. For subject-independent experiments, average accuracy attained 88.37$\%$ for two tasks and a noteworthy 83.03$\%$ for three tasks. These findings highlight the model's superiority over various alternative approaches in emotion recognition.

Furthermore, the authors introduced a visual approach to emphasize the role of EEG channels in emotion recognition. This visualization aids experts in brain science in evaluating result reliability. The visualized outcomes effectively underline key features captured by the DCoT model, shedding light on the significance of individual EEG channels in emotion recognition. This visual interpretability enriches the understanding of the DCoT model's learning process and its application in EEG-based emotion recognition.

In summary, the experiment outcomes using the SEED dataset validate the DCoT model's effectiveness in EEG-based emotion recognition. The inclusion of visualizations strengthens the model's interpretability, enhancing its potential for real-world applications.

## 4.4 Study 4

A novel deep learning model named EEG Conformer was presented by Song et al [1]. that is capable of effectively decoding EEG signals. The authors want to improve interpretability by utilizing visualization techniques to capture both local and global aspects of EEG categorization.

### Data Pre-processing

Initial preprocessing of the raw EEG data is required before it is fed into the model. First, the authors eliminated unnecessary high-frequency and low-frequency noise using band-pass filtering. They preserved task-relevant rhythms by using a 6th-order Chebyshev filter. After that, Z-score normalization was carried out to lessen the data's volatility and non-stationarity. This is how the Z-score normalization was determined [1]:

$$x_o = \frac{x_i - \mu}{\sqrt{\sigma^2}} \tag{4.1}$$

where $x_i$ and $x_o$ stand for the output of standardization and band-pass filtered data, respectively. The mean and variance, denoted by $\mu$ and $\sigma^2$, are computed using the training set of data and applied straight to the test set.

### Network Architecture

The architecture of the EEG conformer network comprises three fundamental components: a convolutional module, a self-attention module, and a fully connected classifier, as shown in Figure 4.6. Each component plays a distinct role in processing preprocessed EEG data and extracting valuable features for accurate decoding.

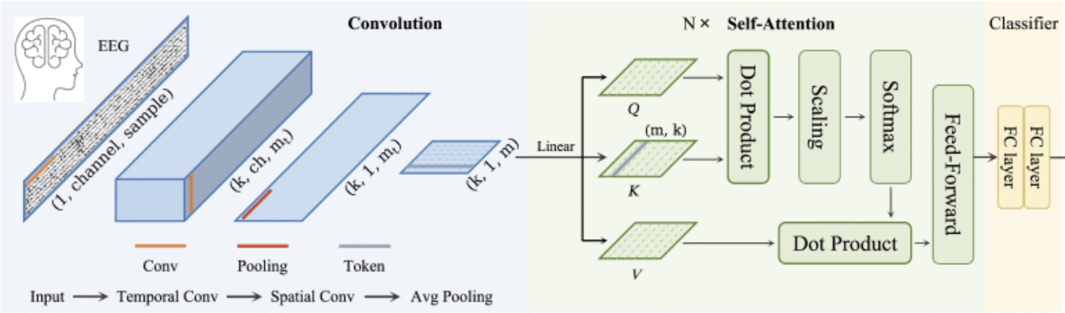
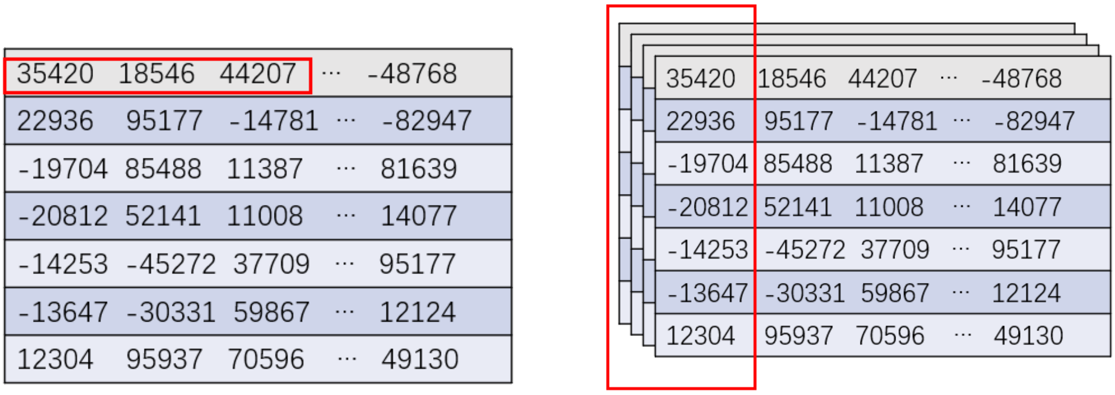


***Figure 4.6.*** *EEG conformer network architecture*

The EEG data comprises several channels, each containing one-dimensional time series data. Prior to inputting this EEG data into the deep learning network, it undergoes a rearrangement and consolidation process.

- **Convolutional module:** A series of preprocessed EEG trials, including the channel and sample dimensions, are provided as input to the convolution module. One dimension is added to each trial to represent the convolution channel. The temporal convolution layer applies a 1D convolution operation along the time dimension of the input, resulting in a tensor with the same number of channels and spatial dimensions but a reduced temporal dimension. The spatial convolution layer applies a 1D convolution operation along the channel dimension of the input, resulting in a tensor with the same number of channels and temporal dimensions but a reduced spatial dimension. The output of the spatial convolution layer is then passed through an average pooling layer, which reduces the spatial dimension further by taking the average of each channel across all spatial positions. The temporal convolution layer (presented in Figure 4.7) is responsible for capturing the temporal correlation present inside each electrode channel. The spatial convolution layer (presented in Figure 4.7) is tasked with capturing the spatial association among distinct electrode channels.



| | | | | |
|---|---|---|---|---|
| 35420 | 18546 | 44207 | ··· | -48768 |
| 22936 | 95177 | -14781 | ··· | -82947 |
| -19704 | 85488 | 11387 | ··· | 81639 |
| -20812 | 52141 | 11008 | ··· | 14077 |
| -14253 | -45272 | 37709 | ··· | 95177 |
| -13647 | -30331 | 59867 | ··· | 12124 |
| 12304 | 95937 | 70596 | ··· | 49130 |

| | | | | |
|---|---|---|---|---|
| 35420 | 18546 | 44207 | ··· | -48768 |
| 22936 | 95177 | -14781 | ··· | -82947 |
| -19704 | 85488 | 11387 | ··· | 81639 |
| -20812 | 52141 | 11008 | ··· | 14077 |
| -14253 | -45272 | 37709 | ··· | 95177 |
| -13647 | -30331 | 59867 | ··· | 12124 |
| 12304 | 95937 | 70596 | ··· | 49130 |

***Figure 4.7.*** *The temporal convolution illustration (left) and the spatial convolution illustration (right). (The data elements seen in the picture do not correspond to those presented in the original text)*

- **Self-attention module:** The primary objective of this module is to extract long-term temporal information by utilizing the feature map that is created by the convolution module, and it serves to enhance the restricted sensory scope that is intrinsic to the convolution module. The output of the convolution module is a tensor with three dimensions: channel, temporal, and spatial. The self-attention module takes this tensor as input and applies a multi-head self-attention mechanism to capture global dependencies between different time positions. The self-attention mechanism involves computing attention scores between each pair of time positions in the input tensor and using these scores to weigh the importance of each time position for each channel. This approach enables the network to effectively highlight significant traits while disregarding inconsequential ones. The method has a high degree of efficacy in capturing intricate nonlinear relationships included in EEG data. Furthermore, the module integrates two fully linked feed-forward layers in order to improve the model's capacity to accurately represent the data. Moreover, it should be noted that the output of the self-attention module is a tensor that possesses an identical form to that of the input tensor.

- **Fully connected classifier:** The output of the self-attention module is a tensor with three dimensions: channel, temporal, and spatial. The fully connected classifier applies a global average pooling operation along the temporal and spatial dimensions of the input tensor, resulting in a tensor with only the channel dimension. This tensor is then passed through two fully connected layers, each followed by a ReLU activation function. The output of the second fully connected layer is a tensor with $M$ dimensions, where $M$ is the number of EEG categories. The softmax function is applied to this tensor to produce a probability distribution over the categories. The entire framework is trained using cross-entropy loss, as defined below:

$$\mathcal{L} = -\frac{1}{N_b} \sum_{i=1}^{N_b} \sum_{c=1}^{M} y \log(\hat{y}) \tag{4.2}$$

where $M$ represents the number of EEG categories, $y$ is the base true label, and $\hat{y}$ is the predicted label. $N_b$ denotes the number of trials in a batch.

### Result

Three distinct publicly accessible datasets were employed in their investigation of EEG conformer, which are BCI Competition IV Dataset 2a [57], BCI Competition IV Dataset 2b [57], SEED dataset [55]:

- **BCI Competition IV Dataset 2a:** The first dataset is the EEG data from 9 individuals from the BCI Competition IV Dataset 2a, which was made available by Graz University of Technology. There were four motor imagery exercises that included the movement of the tongue, both feet and the left and right hands. Twenty-two Ag/AgCl electrode sessions were collected at a sampling rate of 250 Hz on separate days. 288 EEG trials, or 72 trials for each task, were conducted during a single session. They employed [2, 6] seconds for every trial and, following their studies, band-passed the EEG data to [4, 40] Hz. The purpose of the first session was training, and the second was testing.
- **BCI Competition IV Dataset 2b:** The second dataset, which includes EEG data from 9 participants, is the BCI competition dataset 2b from the Graz University of Technology. Two left- and right-handed motor imagery tasks are included in the

dataset. Using three bipolar electrodes (C3, Cz, and C4) and a sampling rate of 250 Hz, the individuals collected data in five sessions, with 120 trials per session. They conducted their experiments using [3, 7] seconds per trial. In addition, we applied band-pass filtering between [4, 40] Hz in order to minimize noise at both high and low frequencies. The training set was used the first three times, and the test set was used the final two times.

- **SEED dataset:** The SEED data set is described in detail in Study 3.

On the first dataset, the EEG conformer model achieves an average classification accuracy of 78.66%, while on the second dataset, it achieves an accuracy of 84.63%. On the third dataset, it can achieve an accuracy of up to 95.3%. They not only calculated the accuracy, but they also counted the inter-rater agreement or reliability of the categorized data as kappa. Kappa is utilized as a metric to assess the overall performance of the model in EEG decoding [1]. It considers both the accuracy of the model and the accuracy of random guessing. On each of the three datasets, the EEG conformer model yields kappa values of 0.7155, 0.6926, and 0.9295 respectively. Overall, the EEG conformer had the best performance on the SEED dataset. This is due to the fact that the SEED dataset was developed specifically for the purpose of emotion recognition, which is a more difficult endeavor than the categorization of motor images, and it requires that the model capture both local and global information contained within the data. The EEG conformer model was developed expressly for the purpose of overcoming this difficulty. It does so by making use of a self-attentive mechanism that discovers long-range relationships and recognizes global patterns in the data. The SEED dataset exceeds the other two in terms of the number of trials and the variety of emotional stimuli it contains.

## 4.5 Discussion of Transformer Network in Analyzing EEG Signal

A noteworthy aspect underscored by the review is the strategic integration of Transformer networks into EEG analysis workflows. The efficacy of these networks lies not only in their inherent capacity to capture complex temporal and spatial relationships within EEG signals but also in their adaptability across diverse cognitive tasks. Particularly relevant is the inclusion of detailed methodologies for data preprocessing, ensuring the input signals are primed for optimal analysis. Techniques such as downsampling, artifact removal, and signal segmentation serve as essential prelude steps that lay the foundation for accurate and meaningful feature extraction. Moreover, the spotlight on Transformer network architectures emphasizes their pivotal role in revolutionizing EEG analysis. Through a combination of self-attention mechanisms, multi-head layers, and feed-forward neural networks, Transformer networks transcend the limitations of conventional methods by effectively capturing dependencies across various frequency bands and channels. In the four literatures reviewed above where transformer architectures have been applied, the data input to the transformer in each model is differentiated:

- **Transformer:** Before sending the EEG data to the transformer encoder, the authors preprocessed it to remove noise and artifacts. In particular, the raw EEG data was imported and band-pass filtering was implemented on the EEG to eliminate environmental and muscle noise. After epoching the data and removing the bad epochs, ICA was applied to remove the bad channels. Embedding and positional coding were performed before sending the input to the encoder. Embedding is the process of converting the input into a fixed-size vector, and position coding is used to provide contextual information about the relative position of the input.
- **TC-Net:** After the data preparation phase, the patch partition module receives the preprocessed EEG signal. The EEG signal of each channel is divided into non-overlapping neighboring patches using the patch partition module in order to maintain the signal's temporal and frequency characteristics. The size of each patch is determined by the frequency and temporal resolution of the input signal. Each segment serves as a marker, summarizing the primary characteristics of the signal. A string of markers is the Patch Partition module's output, and this string is passed into the EEG transformer module.
- **DCoT:** After denoising, segmentation, and feature extraction a feature matrix is obtained, which is then fed into a deep convolutional layer to extract the complete information of the multi-frequency data. To ensure that each channel retains its relative independence, the DCoT model partitions the EEG feature matrix into a sequence of one-dimensional vectors corresponding to the EEG channels, and each vector represents the EEG features of one channel. These vectors are then fed into the transformer encoder, which processes the data from all channels in parallel. In order to preserve position information, a one-dimensional learnable position embedding is appended to the input sequence to the encoder in the transformer.
- **EEG conformer:** A preprocessed EEG with a convolutional channel that has one dimension added to both the channel and the sample. The convolution module then extracts local temporal and spatial properties from the preprocessed EEG data. To capture temporal relationships, temporal convolution was used along the temporal dimension; to capture spatial dependencies, spatial convolution was applied along the electrode channels. The self-attention module of the converter module then receives the generated spatial-temporal representation.

In summary, transformer networks have transformative potential in EEG signal analysis, with advantages in capturing intricate temporal patterns and greater adaptability to different cognitive tasks. It is foreseeable that as the research community continues to embrace and advance this innovative paradigm, the integration of transformer networks will not only refine the understanding of EEG data, but also provide new avenues for groundbreaking applications in neuroscience, clinical diagnostics, and human-computer interfaces.

# 5. TRANSFORMER IN THE ASSESSMENT OF COGNITIVE LOAD: CASE STUDY

Cognitive load, the mental effort required to process information, is a fundamental concept in cognitive psychology and human-computer interaction. Understanding cognitive load is essential in designing effective learning environments, optimizing user interfaces, and enhancing overall cognitive performance. As technology continues to evolve and infiltrate various aspects of our lives, the need for accurate and efficient methods to assess cognitive load becomes increasingly vital. This chapter delves into a case study that explores the application of transformer models in the assessment of cognitive load. Transformer models have revolutionized the field of natural language processing (NLP) and have shown promise in various applications across domains. Their ability to capture complex dependencies in sequential data has made them a compelling choice for analyzing cognitive processes.

The primary objective of this case study is to explore the application of the EEG conformer [1] network mentioned in Chapter 4 in assessing cognitive load. This chapter is structured into several sections to provide a comprehensive overview of the case study: data collection, data pre-processing, network architecture, and results.

## 5.1 Data Collection

In this case study, two distinct datasets formed the basis for our exploration of cognitive load assessment. The initial dataset, known as the Simultaneous Task EEG Workload (STEW) dataset [46], encompasses a diverse collection of raw EEG data acquired from participants who were engaged in a multitasking workload experiment, a dataset also employed in Study 1 of the literature review. Furthermore, as part of the comprehensive description of the STEW dataset outlined in Study 1, each participant contributed 2.5 minutes of EEG recordings during both resting-state conditions and multitasking scenarios with the SIMKAP test. During these multitasking phases, participants were requested to provide self-assessments of their perceived cognitive workload, utilizing a rating scale ranging from 1 to 9. These subjective ratings offered valuable insights into their cognitive load experiences.

The second dataset originated locally from a cohort of 30 individuals aged between 18 and 65, all devoid of any psychiatric history. These participants actively engaged in the N-back memory game during data collection, a task involving the display of single digits (ranging from 0 to 9) on a laptop screen every 3 seconds. Their objective was to click the mouse when a specific target digit appeared. The game consisted of 9 rounds, with brief pauses, accumulating to 90 seconds per round. In level 0, participants were instructed to click upon sighting the predetermined target number. In level 1, the instruction was to click if the previously shown number reappeared, and in level 2, participants responded when the number from two steps back was displayed. Data acquisition was executed using the Neuroelectrics® Instrument Controller (NIC2) software, in tandem with the ENOBIO® EEG recording system supplied by Neuroelectrics®. EEG recordings were performed across 7 channels, with a primary focus on the prefrontal region, sampled at a frequency of 500 Hz [58].

## 5.2 Data Pre-processing

Within the data preprocessing framework, a uniform methodology was consistently applied to the STEW dataset and locally acquired data, ensuring standardized data preparation for subsequent analysis. Initially, the code is designed to extract raw EEG data from their respective source files, transforming them into a digital array format suitable for further processing. Simultaneously, an array is initialized as a container for EEG data refinement and denoising.

The processing of individual EEG channels is given careful attention as the code develops. In this phase, the important frequency components of the EEG signal are retained while unwanted noise components are filtered out using a precise bandpass Butterworth filter. The code then explores the crucial discrete wavelet transform (DWT) stage. Because of their similarity to blinking patterns, the use of Daubechies 4 (db4) mother wavelets is especially noteworthy in this instance. Through this complex procedure, the EEG signal is broken down into its fundamental frequency components, revealing the underlying frequency patterns and subtleties.

Furthermore, the algorithm calculates necessary thresholds, which is a critical step in the denoising procedure. These threshold values, derived from the median absolute value of wavelet coefficients, serve as a discriminative mechanism to separate valuable neural signals from extraneous noise. The code employs the $cmddenoise$ function, a trusted tool for noise reduction, to meticulously eliminate noise and unwanted artifacts from the EEG signals. This critical step contributes significantly to refining the EEG signal dataset, fine-tuning the data by ridding it of unwanted elements while retaining the essential neural information, see the comparison from Figure 5.1.

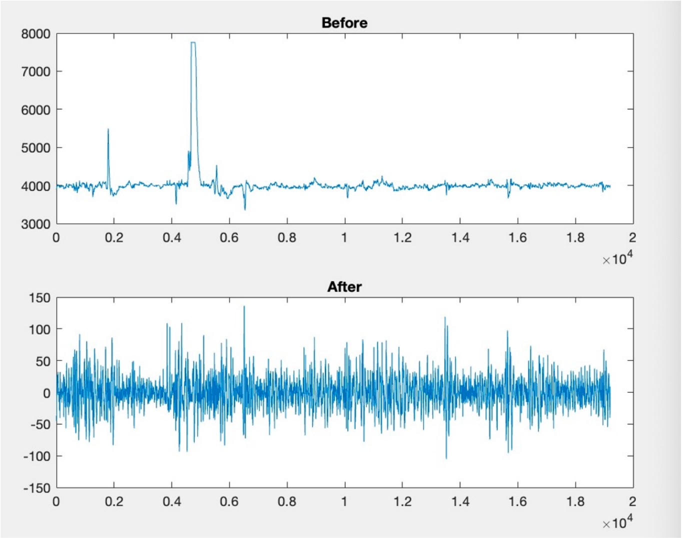


***Figure 5.1.*** *The comparison of EEG signal before pre-processing and after pre-processing*

## 5.3 Implementation

In the case study, implementing this deep learning neural network involves dealing with two datasets: the local dataset and the STEW dataset. The official repository of EEG Conformer [59] is used as a reference in the design and implementation of validation experiments.

The following lists the software and hardware environments used by the implementation. The software environments include the programming language used and the names and version numbers of the major libraries. The hardware environment includes the GPU, RAM, and other information.

Software Environment:

1. Programming Language: Python 3.10.0
2. Deep Learning Packages: PyTorch 2.1.0, Pandas 2.1.1, SciPy 1.11.3

Hardware Environment: Google Colab (GPU Type: V100; GPU RAM: 15GB; System RAM: 40GB).

The program is thoughtfully designed with two primary considerations:

- **Modularization of the deep learning neural network:** Modularity is a key feature of the design, allowing for ease of future modifications and structural adjustments to the model. As shown in Figure 5.2, the implementation of this deep learning neural network involves four essential blocks, each associated with its specific input variables, parameters, and output variables. This modular structure not only enables the seamless integration of additional layers to the MLP (Multi-Layer Perceptron) but also simplifies the process of making program extensions or modifications, such as altering the convolution technique.

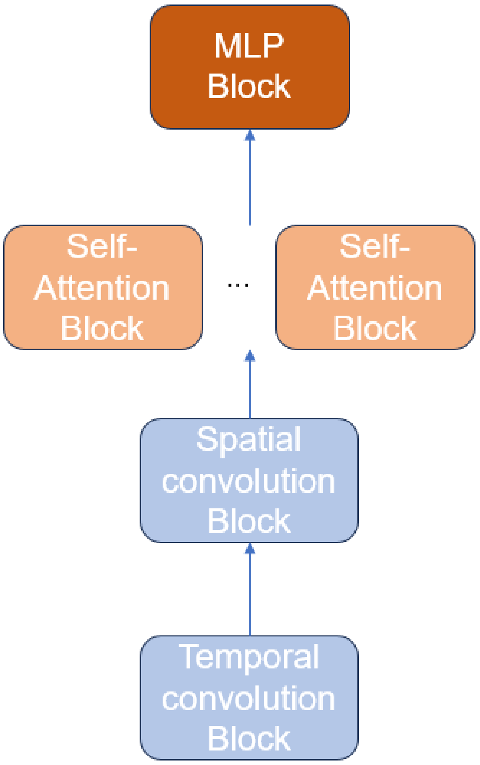


***Figure 5.2.*** *The flow of the program module*

- **Parameterization of the modules:** Each module within the network has its own set of parameters. This parameterization empowers the user to fine-tune the neural network, enabling the exploration of its performance across different configurations. Within deep learning neural networks, various tunable parameters exist, and adjusting these parameters is pivotal in optimizing the network's performance across different datasets. Furthermore, this parameterization serves as a valuable tool for conducting experiments and thoroughly assessing the neural network's capabilities and adaptability. Table 5.1 presents several critical parameters within various modules that significantly impact the neural network's performance, as elaborated upon in the subsequent section.

  The configuration of the convolution kernel affects the size of the token that the convolution layer processes within the convolution module, and the number of channels is dataset-dependent. The three parameters within the multi-head attention module significantly impact the computation vector and overall performance. Additionally, the parameters of the MLP classifier are contingent on the number of classifications required.

***Table 5.1.*** *Parameters of modules*

| | |
|---|---|
| Convolution Module | channel number |
| | kernel size |
| Multi-head Attention Module | number of Heads |
| | depth |
| | embedding size |
| MLP Classifier | number of classes |

## 5.4 Result

This part is devoted to the presentation and comprehensive assessment of the performance of the EEG conformer network application on two separate datasets: the STEW dataset and a dataset obtained locally. By systematically manipulating a variety of parameter combinations, the primary objective of this study is to assess the robustness and effectiveness of the model across a number of experimental scenarios. These configurations include the number of heads in the self-attention mechanism, the depth of the architecture, and the size of the convolution kernel of the convolution module. This allows us to understand how their model performance is affected by changing these parameters.

### 5.4.1 STEW Dataset

- **The number of heads:** In transformer-based deep learning neural networks, the number of heads significantly influences their capacity to identify and extract global features and patterns from the data. As illustrated in Figure 5.3, when the number of heads is less than 8, the accuracy rate increases as the number of heads grows. With fewer than 8 heads, augmenting the number of heads effectively enhances the neural network's learning capability, achieving peak accuracy in the experiment when there are 8 heads. However, when the number of heads exceeds 8, the accuracy rate demonstrates a declining trend. This phenomenon arises because, as the number of heads increases, the neural network's demand for data volume also rises. Therefore, using more than 8 heads can lead to overfitting when data is insufficient, causing a decrease in the accuracy rate.

  Hence, there exists a trade-off when deciding on the number of heads:

  1. **As the number of heads increases, the amount of data required also grows proportionally.** The self-attention mechanism inherently demands a larger volume of data, and when the number of heads is increased excessively, it becomes more susceptible to overfitting, thereby leading to a reduction in accuracy.

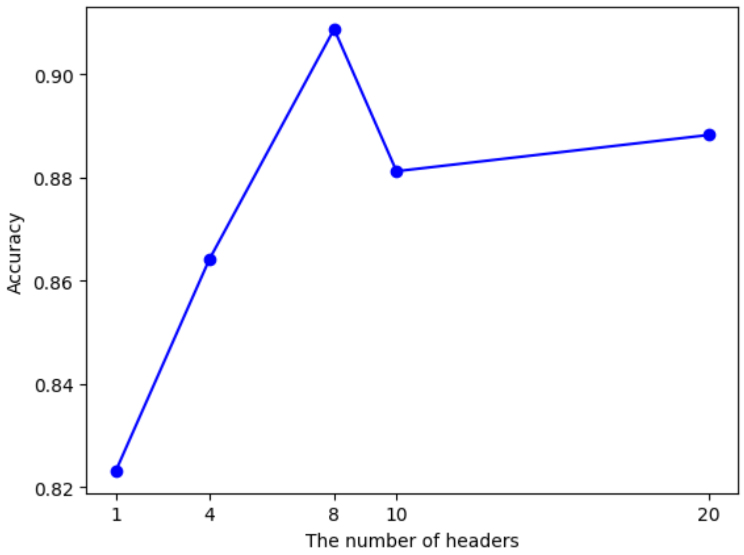


***Figure 5.3.*** *The effect of the number of heads in the self-attention module on accuracy.*

2. **Increasing the number of heads also escalates the computational cost.** With an increase in the number of heads, there is an accompanying rise in the memory space needed to perform gradient computations, and simultaneously, the time required for self-attention computations also increases.

- **Depth:** The depth of the self-attention module and the number of heads have similar effects on accuracy and computational cost. In general, increasing the depth of the self-attentive module and adding more layers improves its ability to extract and learn information, but it also comes at the cost of increased time and memory requirements for training and reasoning. Figure 5.4 illustrates that the highest accuracy is attained at a depth of 3.

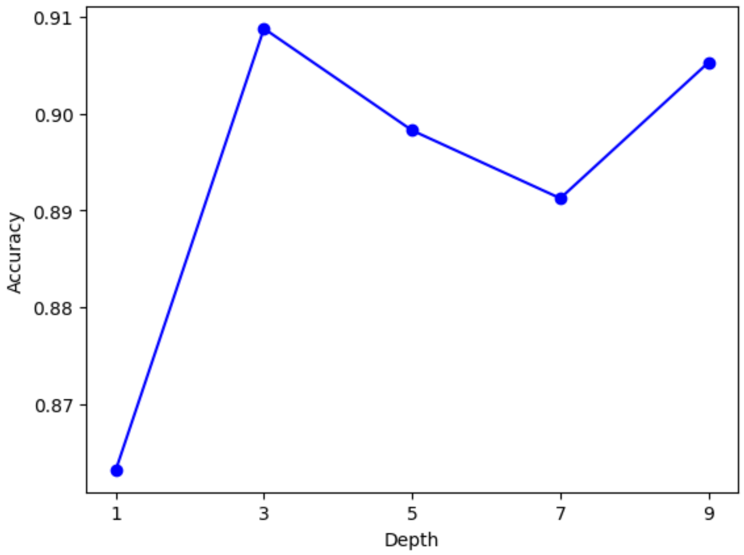


***Figure 5.4.*** *The effect of depth in the self-attention module on accuracy.*

- **The size of convolution kernel:** In the convolution module, the kernel size, often referred to as the filter size, plays a crucial role in determining the scale of information captured. Smaller kernels excel at capturing localized information, whereas larger kernels are more effective at capturing global features. As previously discussed, in EEG Conformer, the convolution module focuses on extracting localized information, while the self-attention module handles global information extraction. The input to the self-attention module is derived from the local feature map produced by the convolution module. Figure 5.5 illustrates the relationship between kernel size and accuracy, showing that accuracy increases as the local field of view expands and reaches its peak at a kernel size of 30.

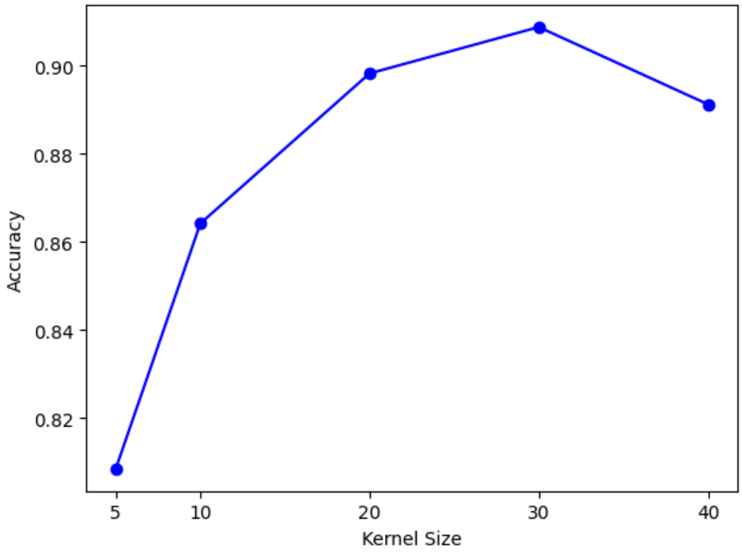


***Figure 5.5.*** *The effect of the size of the convolution kernel in the convolution module on accuracy.*

  Regarding computational cost, smaller kernels result in larger feature maps, while larger kernels produce smaller feature maps. Additionally, using large kernels may increase the risk of overfitting.

- **Convergence analysis:** As depicted in Figure 5.6, the optimal experimental outcome for the STEW dataset was attained with the following parameter settings: 4 attention heads, a kernel size of 30, and a depth of 3, resulting in an accuracy of 90.8772$\%$. The training process converged after 125 epochs.

### 5.4.2 Local Dataset

Our series of parametric experiments on the locally gathered dataset showed that varying the model's depth and number of attention heads had comparable impacts to those we had previously shown in the STEW dataset. The observed consistent patterns indicate

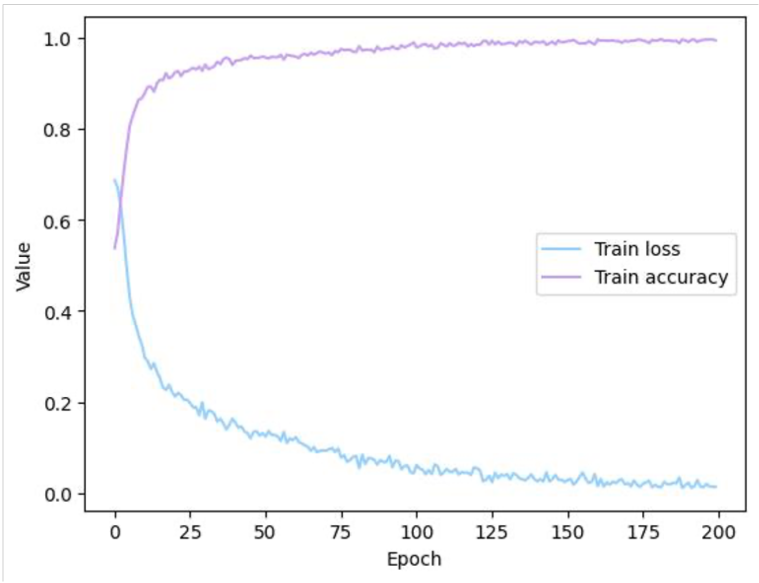


***Figure 5.6.*** *The convergence illustration of accuracy.*

that the EEG conformer network's reaction to variations in the number of attention heads and its architectural complexity displays a level of universality across diverse datasets.

Nevertheless, it is crucial to emphasize a significant disparity between the two datasets. The mismatch became apparent when examining the effects of altering the kernel size in the convolution module of the EEG conformer network. In contrast to the constant patterns observed in attention heads and depth, the impact of kernel size on the experimental outcomes was shown to be different and dependent on the specific dataset.

This dataset-dependent behavior underscores the importance of considering dataset characteristics and domain-specific nuances when optimizing the model's hyperparameters.

- **The number of heads:** The examination of experimental findings reveals a notable discrepancy between the local dataset and the STEW dataset, particularly in their responsiveness to changes in the number of attention heads utilized in the EEG conformer network. The graphical representation in Figure 5.7 clearly demonstrates that the local dataset has a greater sensitivity to variations in the number of attention heads compared to the STEW dataset. It is noteworthy to mention that the accuracy measure displays a discernible pattern in relation to the quantity of heads.

  In the context of the local dataset, it is apparent that the number of attention heads and the performance of the model are positively correlated in the context of the local dataset. The degree of ascension is notably prominent in cases where the quantity of headings is less than three. This implies that the EEG conformer network can derive advantages from enhanced attention granularity, enabling it to capture more nuanced linkages and dependencies within the complex EEG recordings of

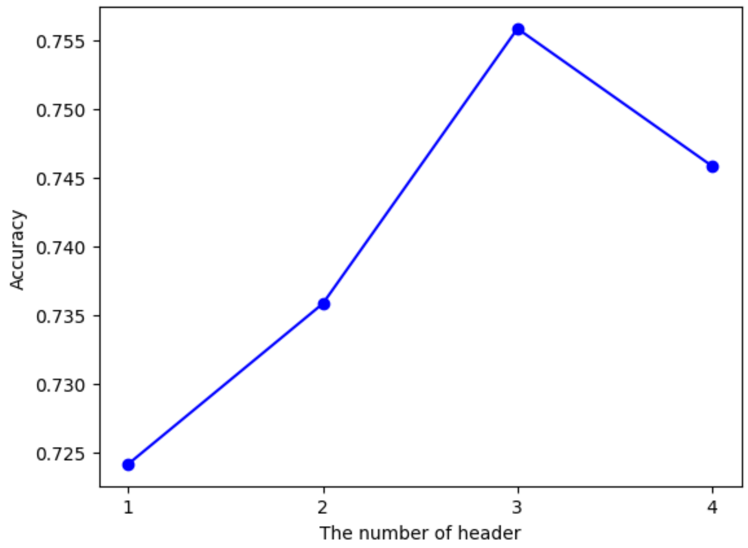


***Figure 5.7.*** *The effect of the number of heads in the self-attention module on accuracy.*

the local dataset. However, a critical juncture arises when the quantity of attention headings approaches four. At this point in time, there is a noticeable decrease in the accuracy metric. The observed decline in performance is indicative of a situation that is frequently associated with overfitting, a condition in which the model becomes overly customized to the training set, impairing its capacity to predict new, unseen data with any degree of accuracy. The intricacy of the dataset at hand seems to require a larger quantity of data in order to adequately mitigate the problem of overfitting.

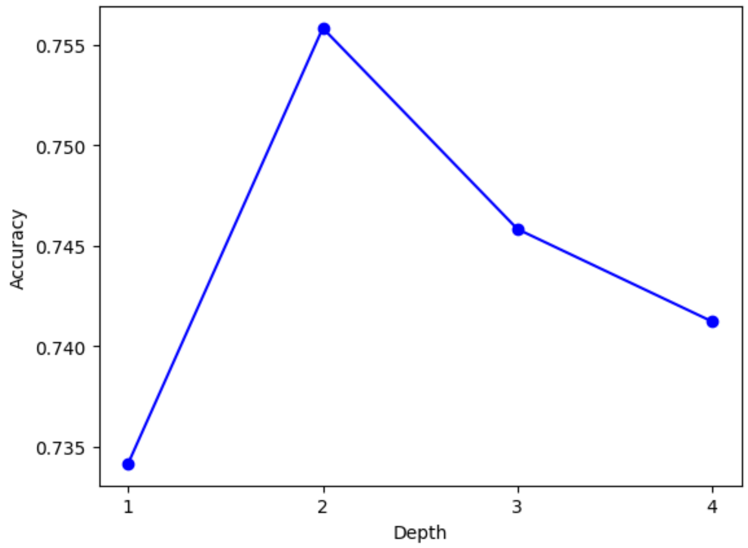


***Figure 5.8.*** *The effect of depth in the self-attention module on accuracy.*

- **Depth:** Our experimental investigation on the impact of the depth of the self-attention module in the EEG Conformer network has revealed a significant pattern that offers vital insights into the model's interaction with the specific details of the local dataset. To be more precise, the model showed overfitting when we increased the depth over a threshold that we found to be 2. This phenomenon resulted in a noticeable decrease in accuracy.

By augmenting the depth of the self-attention module, the model's ability to effectively capture and represent the patterns and relationships within the training data was enhanced. Nevertheless, this increased capability came with a trade-off, as the model progressively improved its ability to capture the intricate particulars and subtleties of the training data, encompassing the irrelevant variations and peculiarities that might not effectively apply to novel, unfamiliar data. One of the outcomes of overfitting is a significant decrease in accuracy. When the model is confronted with data that exhibits even minor deviations from the training set, it experiences difficulties in generating precise predictions. This is due to the model's tendency to essentially "memorize" the training data rather than acquire a comprehensive understanding of universal patterns and relationships.

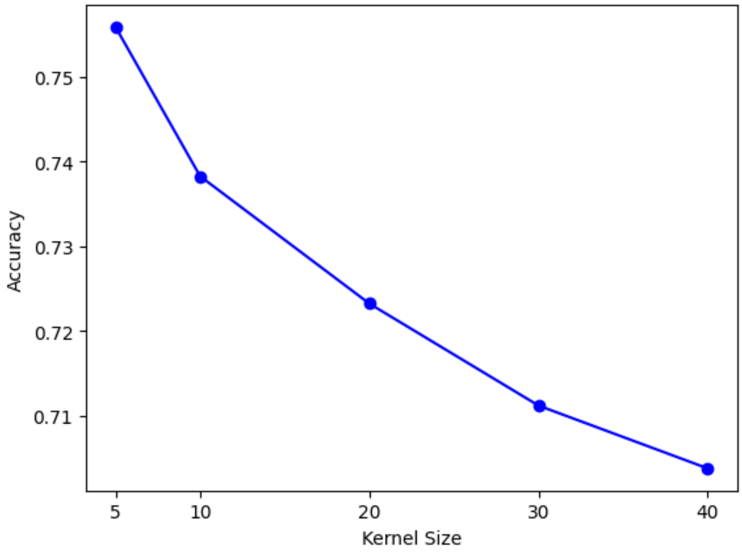


***Figure 5.9.*** *The effect of the size of the convolution kernel in the convolution module on accuracy.*

- **The size of convolution kernel:** As depicted in Figure 5.9, a discernible trend emerges wherein the accuracy experiences a gradual decline as the kernel size increases. There are two primary explanations that can account for this observed decrease in accuracy:

  1. As the size of the kernel increases, there is an increased likelihood of encountering overfitting in the model. In the context of kernel size, a larger kernel possesses an increased ability to catch intricate information that exists in the training data, such as noise. However, this heightened capacity may not effectively generalize to unseen data.

2. The use of a large kernel results in a lack of localized information in the extracted feature maps, which may not provide appropriate input to the subsequent self-attention module.

To summarize, the local dataset achieved the highest accuracy: 75.5827$\%$ with a convolution module kernel size of 5, a self-attention module depth of 2, and a head number of 3.

# 6. DISCUSSION AND CONCLUSIONS

The use of transformer neural networks for the processing of EEG signals is a rapidly developing and potentially game-changing topic that lies at the crossroads of deep learning and cognitive research. The EEG, which is a very effective technique for measuring the activity of the brain in real-time, offers a glimpse into the workings of the cognitive process. Therefore, this thesis proceeds from a description of transformer network designs and cognitive load to investigate 4 transformer models, their combinations with other models in EEG analysis, and a case study of one of them. Therefore, this thesis first describes the theories and measurements related to EEG-based cognitive load and then explains the basic architecture and operation of transformer networks. Subsequently, several transformer models and their combined models with other network architectures in EEG analysis are explored, and a case study of one of the models is conducted. Throughout the learning and experimentation process, the following points are crucial in determining the success of the experiment:

- **The quality of the data:** High-quality data ensures that the insights gained from the EEG signal accurately represent neural processes, which is crucial for making a correct diagnosis or understanding cognitive patterns. In addition, transformer-based models require high-quality data for generalization, just like any machine learning model. And the quality of the data depends on upfront data collection and proper data preprocessing. During the data collection process, it is important to avoid, as much as possible, any factors that can be completely avoided to affect the data. In data preprocessing, important information should not be lost while removing noise and artifacts, and preprocessing techniques should be selected based on the data.
- **Selection of model:** The study objectives, data features, available data, and problem complexity need to be considered when choosing an appropriate model for EEG analysis. There is also a need to review the existing literature in the field of EEG analysis to see which models have been successfully applied to similar tasks, which can provide insight into the effectiveness of different approaches. Additionally, it is helpful to experiment with multiple models to see which one performs best on a particular EEG dataset and task. Model performance is often evaluated using appropriate evaluation metrics (such as accuracy, F1 score, or kappa). In some

cases, combining multiple models into a whole can improve performance and robustness.

- **Parameter tuning:** Tuning the parameters of the model applied in EEG analysis is a critical step in optimizing its performance. This process includes techniques such as understanding the parameters of the model, defining the parameter search space, and performing cross-validation to evaluate different settings for optimization. Furthermore, continuous iteration and refinement of the parameters, as well as the consideration of regularization techniques are key. Visualization of results and maintenance of records are also important for tracking progress. Overall, parameter tuning is an iterative process that ensures optimal performance of the model for EEG analysis tasks.
- **Computational infrastructure:** Computational resources are critical in EEG analysis studies involving the application of transformer models. Transformers are complex deep-learning architectures, and they require a significant amount of computational power to accomplish a variety of tasks. Computational infrastructure facilitates efficient research, collaboration, and experimentation in EEG analysis, enabling researchers to discover valuable insights.

While transformer networks and network architectures that include other networks have made notable progress and gained substantial benefits in this area, there remain some difficulties that require future attention and resolution:

- **Computational resources:** Computational effort may be required for transformer models, particularly when working with big EEG datasets. It is possible that the training and deployment of these models may demand large computational resources, which may be a hindrance in certain research contexts.
- **Data requirements:** In order to train transformers effectively, it is often necessary to have access to substantial annotated datasets. The process of acquiring and categorizing EEG data, particularly when accompanied by accurate cognitive load labels, might prove to be a laborious and costly endeavor.
- **Model complexity:** The complexity of transformer architectures can pose challenges for researchers and practitioners who are not familiar with deep learning techniques. Implementation and fine-tuning of transformer models can be non-trivial.
- **Overfitting:** Transformers are prone to overfitting, especially when trained on limited data. Careful regularization and validation techniques are necessary to prevent overfitting and ensure generalizability.

In summary, transformer networks have emerged as a potent and adaptable tool in the realm of EEG research, providing vital insights and progress in diverse disciplines. These networks have been utilized in several domains such as cognitive load assessment, the study of functional connectivity, recognition of emotions, research of sleep patterns, development of brain-computer interfaces, and other related applications. The capacity to effectively represent intricate temporal relationships, dynamically extract distinctive characteristics, and offer immediate insights has significantly broadened the scope of EEG research and its practical implementations. However, challenges remain. Future developments will likely involve addressing these challenges, exploring more efficient architectures, and improving data acquisition techniques. Despite these challenges, the transformative potential of transformer networks in EEG analysis is undeniable, and their continued integration promises to advance the understanding of cognitive processes and brain activity, thereby enhancing diagnostic and therapeutic applications in the fields of neurology, psychology, and human-computer interaction.

# REFERENCES

[1] Song, Y., Zheng, Q., Liu, B. and Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 31 (2022), pp. 710–719.

[2] Haas, L. F. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry* 74.1 (2003), pp. 9–9.

[3] Teplan, M. et al. Fundamentals of EEG measurement. *Measurement science review* 2.2 (2002), pp. 1–11.

[4] Tudor, M., Tudor, L. and Tudor, K. I. Hans Berger (1873-1941)–the history of electroencephalography. *Acta medica Croatica: casopis Hravatske akademije medicinskih znanosti* 59.4 (2005), pp. 307–313.

[5] Goverdovsky, V., Looney, D., Kidmose, P. and Mandic, D. P. In-Ear EEG From Viscoelastic Generic Earpieces: Robust and Unobtrusive 24/7 Monitoring. *IEEE Sensors Journal* 16.1 (2016), pp. 271–277. DOI: 10.1109/JSEN.2015.2471183.

[6] Kidmose, P., Looney, D., Ungstrup, M., Rank, M. L. and Mandic, D. P. A Study of Evoked Potentials From Ear-EEG. *IEEE Transactions on Biomedical Engineering* 60.10 (2013), pp. 2824–2830. DOI: 10.1109/TBME.2013.2264956.

[7] Arquilla, K., Webb, A. K. and Anderson, A. P. Textile electrocardiogram (ECG) electrodes for wearable health monitoring. *Sensors* 20.4 (2020), p. 1013.

[8] Li, G.-L., Wu, J.-T., Xia, Y.-H., He, Q.-G. and Jin, H.-G. Review of semi-dry electrodes for EEG recording. *Journal of Neural Engineering* 17.5 (2020), p. 051004.

[9] Mecarelli, O. Electrode placement systems and montages. *Clinical Electroencephalography* (2019), pp. 35–52.

[10] Jasper, H. H. Ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.* 10 (1958), pp. 371–375.

[11] Rojas, G. M., Alvarez, C., Montoya, C. E., De la Iglesia-Vaya, M., Cisternas, J. E. and Gálvez, M. Study of resting-state functional connectivity networks using EEG electrodes position as seed. *Frontiers in neuroscience* 12 (2018), p. 235.

[12] Klem, G. H. The ten-twenty electrode system of the international federation. The international federation of clinical neurophysiology. *Electroencephalogr. Clin. Neurophysiol. Suppl.* 52 (1999), pp. 3–6.

[13] Seeck, M., Koessler, L., Bast, T., Leijten, F., Michel, C., Baumgartner, C., He, B. and Beniczky, S. The standardized EEG electrode array of the IFCN. *Clinical neurophysiology* 128.10 (2017), pp. 2070–2077.

[14] Zeynali, M. and Seyedarabi, H. EEG-based single-channel authentication systems with optimum electrode placement for different mental activities. *biomedical journal* 42.4 (2019), pp. 261–267.

[15] Nayak, C. S. and Anilkumar, A. C. *EEG Normal Waveforms. StatPearls*. 2020.

[16] Cannon, J., McCarthy, M. M., Lee, S., Lee, J., Börgers, C., Whittington, M. A. and Kopell, N. Neurosystems: brain rhythms and cognitive processing. *European Journal of Neuroscience* 39.5 (2014), pp. 705–719.

[17] Watson, B. O. and Buzsáki, G. Sleep, memory & brain rhythms. *Daedalus* 144.1 (2015), pp. 67–82.

[18] Urigüen, J. A. and Garcia-Zapirain, B. EEG artifact removal—state-of-the-art and guidelines. *Journal of neural engineering* 12.3 (2015), p. 031001.

[19] Sörnmo, L. and Laguna, P. *Bioelectrical signal processing in cardiac and neurological applications*. Vol. 8. Academic press, 2005.

[20] Clark, B. R. *Creating entrepreneurial universities: organizational pathways of transformation. Issues in Higher Education.* ERIC, 1998.

[21] Jiang, X., Bian, G.-B. and Tian, Z. Removal of artifacts from EEG signals: a review. *Sensors* 19.5 (2019), p. 987.

[22] Rashmi, C. and Shantala, C. EEG artifacts detection and removal techniques for brain computer interface applications: A systematic review. *Int. J. Adv. Technol. Eng. Explor* 9 (2022), p. 354.

[23] Kaya, I. A brief summary of EEG artifact handling. *Brain-computer interface* 9 (2019).

[24] Mumtaz, W., Rasheed, S. and Irfan, A. Review of challenges associated with the EEG artifact removal methods. *Biomedical Signal Processing and Control* 68 (2021), p. 102741.

[25] Nolan, H., Whelan, R. and Reilly, R. B. FASTER: fully automated statistical thresholding for EEG artifact rejection. *Journal of neuroscience methods* 192.1 (2010), pp. 152–162.

[26] Debie, E., Rojas, R. F., Fidock, J., Barlow, M., Kasmarik, K., Anavatti, S., Garratt, M. and Abbass, H. A. Multimodal fusion for objective assessment of cognitive workload: a review. *IEEE transactions on cybernetics* 51.3 (2019), pp. 1542–1555.

[27] Moray, N. *Mental workload: Its theory and measurement*. Vol. 8. Springer Science & Business Media, 2013.

[28] Wickens, C. D. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3.2 (2002), pp. 159–177.

[29] Cain, B. A review of the mental workload literature. (2007).

[30] Abbass, H. A., Mount, W. M., Tucek, D. and Pinheiro, J.-P. Towards a code of best practice for evaluating air traffic control interfaces. *Australian Transport Research Forum, Adelaide, Australia*. 2011.

[31] Sweller, J. Cognitive load theory. *Psychology of learning and motivation*. Vol. 55. Elsevier, 2011, pp. 37–76.

[32] Plass, J. L., Moreno, R. and Brünken, R. Cognitive load theory. (2010).

[33] Ericsson, K. A. and Kintsch, W. Long-term working memory. *Psychological review* 102.2 (1995), p. 211.

[34] Heard, J., Harriott, C. E. and Adams, J. A. A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems* 48.5 (2018), pp. 434–451.

[35] Hart, S. G. and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

[36] Reid, G. B. and Nygren, T. E. The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 185–218.

[37] Hirshfield, L. M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E. T., Jacob, R. J., Sassaroli, A. and Fantini, S. Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: 5th International Conference, FAC 2009 Held as Part of HCI International 2009 San Diego, CA, USA, July 19-24, 2009 Proceedings 5*. Springer Berlin Heidelberg. 2009, pp. 239–247.

[38] Nie, Y., Tong, S., Li, J., Zhang, Y., Zheng, C. and Fan, B. Time identification of design knowledge push based on cognitive load measurement. *Advanced Engineering Informatics* 54 (2022), p. 101783.

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[40] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. *Attention is all you need*. 2017. URL: http://www.aiotlab.org/teaching/intro2ai/slides/10_attention_n_bert.pdf.

[41] Lai, S., Liu, K., He, S. and Zhao, J. How to generate a good word embedding. *IEEE Intelligent Systems* 31.6 (2016), pp. 5–14.

[42] Ba, J. L., Kiros, J. R. and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[43] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. and Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[44] Lin, T., Wang, Y., Liu, X. and Qiu, X. A survey of transformers. *AI Open* (2022).

[45] Siddhad, G., Gupta, A., Dogra, D. P. and Roy, P. P. Efficacy of transformer networks for classification of raw EEG data. *arXiv preprint arXiv:2202.05170* (2022).

[46] Lim, W., Sourina, O. and Wang, L. P. STEW: Simultaneous task EEG workload data set. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 26.11 (2018), pp. 2106–2114.

[47] Bratfisch, O. and Hagman, E. Simkap–simultankapazität/multi-tasking. *Mödling: Schuhfried GmbH* (2008).

[48] Chakladar, D. D., Dey, S., Roy, P. P. and Dogra, D. P. EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm. *Biomedical Signal Processing and Control* 60 (2020), p. 101989.

[49] Zhu, T., Luo, W. and Yu, F. Convolution-and attention-based neural network for automated sleep stage classification. *International Journal of Environmental Research and Public Health* 17.11 (2020), p. 4152.

[50] Zheng, X. and Chen, W. An attention-based bi-LSTM method for visual object classification via EEG. *Biomedical Signal Processing and Control* 63 (2021), p. 102174.

[51] Wei, Y., Liu, Y., Li, C., Cheng, J., Song, R. and Chen, X. TC-Net: A Transformer Capsule Network for EEG-based emotion recognition. *Computers in Biology and Medicine* 152 (2023), p. 106463.

[52] Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A. and Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

[53] Katsigiannis, S. and Ramzan, N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE journal of biomedical and health informatics* 22.1 (2017), pp. 98–107.

[54] Guo, J.-Y., Cai, Q., An, J.-P., Chen, P.-Y., Ma, C., Wan, J.-H. and Gao, Z.-K. A Transformer based neural network for emotion recognition and visualizations of crucial EEG channels. *Physica A: Statistical Mechanics and its Applications* 603 (2022), p. 127700.

[55] Zheng, W.-L. and Lu, B.-L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on autonomous mental development* 7.3 (2015), pp. 162–175.

[56] Duan, R.-N., Zhu, J.-Y. and Lu, B.-L. Differential entropy feature for EEG-based emotion classification. *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2013, pp. 81–84.

[57] *BCI Competition IV*. URL: `https://www.bbci.de/competition/iv/`.

[58] Beiramvand, M., Lipping, T., Karttunen, N. and Koivula, R. Mental Workload Assessment using Low-Channel Prefrontal EEG Signals. *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*. IEEE. 2023, pp. 1–5.

[59] Song, Y., Zheng, Q., Liu, B. and Gao, X. *EEG Conformer Github*. URL: `https://github.com/eeyhsong/EEG-Conformer.git`.